0-315-01249-8

National Library of Canada
Canadian Theses Division
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Division des thèses canadiennes

49107

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

• Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

MARY ELIZABETH SHIELDS.

Date of Birth — Date de naissance

30 JUNE 1947

Country of Birth — Lieu de naissance

CANADA.

Permanent Address — Résidence fixe

18405 62B AVENUE, EDMONTON, ALBERTA
T5T 2S9.

Title of Thesis — Titre de la thèse

An Item Analysis Method for Evaluating Competency Based Test Items.

University — Université

University of Alberta

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

M. Education

Year this degree conferred — Année d'obtention de ce grade

1980

Name of Supervisor — Nom du directeur de thèse

Dr. Milt Petruk

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date

Oct 15, 1980

Signature

Mary E. Shields

NL-91 (4/77)

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

Ottawa, Canada
K1A 0N4

NL-339 (Rev. 8/80)

THE UNIVERSITY OF ALBERTA

AN ITEM ANALYSIS METHOD FOR EVALUATING COMPETENCY BASED TEST ITEMS

by

© MARY ELIZABETH SHIELDS

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF EDUCATION

IN

VOCATIONAL EDUCATION

DEPARTMENT Industrial and Vocational Education

EDMONTON, ALBERTA

FALL, 1980

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled An Item Analysis Method For Evaluating Competency Based Test Items submitted by Mary Elizabeth Shields in partial fulfilment of the requirements for the degree of Master of Education in Vocational Education.

.....
Supervisor

.....

.....

.....

Date September 4, 1980

Abstract

A 128 item examination was produced by randomly selecting items from an existing item bank at the University of Alberta Hospital School of Nursing. The items were evaluated as to the amount of information they provided in differentiating between 62 nursing students who were within one month of graduating from diploma granting nursing schools in Alberta, and 89 beginning students who were within one month of starting courses at the same schools.

Classical item analysis techniques were adapted in such a way as to compare both groups on several criteria. The resultant item analysis data, consisting of difficulty level, point biserial correlation and distractor discriminating power, was compared to previously computed ideal values for difficulty levels, point biserial correlations and distractor discriminating powers.

The 128 test items were evaluated using three different analysis methodologies in order to determine which method produced the most useful information for item revisions. The three analysis methods which were compared were a classical test theory analysis, a comparison matrix of the beginning and graduating students, and the adapted analysis methodology described above. Suggestions for types of revisions needed for items were given according to the information provided through the adapted item analysis method.

TABLE OF CONTENTS

CHAPTER		PAGE
I	INTRODUCTION	1
	Statement of the Problem	4
	Need for the Study	5
	Delimitations	6
	Limitations	7
	Assumptions	7
	Questions to be Answered	8
	Definition of Terms	9
	Summary	13
II	REVIEW OF RELATED LITERATURE	14
	Uses and Structure	14
	Norm and criterion referenced testing	15
	Formulation of Tests	16
	Relationship to course objectives	17
	Item reliability	19
	Interpretation of Test Scores	20
	True score	21
	Validity	23
	Reliability	25
	Item Analysis	26
	Classical test theory	27
	Latent trait model	29
	Item banking	29
	Summary	30
III	METHODOLOGY AND INSTRUMENTATION	31
	Development of Item Bank	32
	Development of Instrument	34
	Administration of Instrument	35
	Examination Analysis	36
	Item analysis	36
	Comparison of means	37

Adapted Item Evaluation Methodology 37
Computation of Item Analysis Values for Ideal Items 39
Four distractor items 40
Five distractor items 41
Comparative analysis 41
Comparative Assessment of Individual Items 43
Summary 43

IV ANALYSIS OF RESULTS 46
Test Analysis Data Results 46
Classical Analysis 47
Beginner - Graduate Difficulty Level Matrix 49
Comparison of matrix sorting to classical analysis 54
Comparison of Actual and Ideal Performance for an Item 54
Comparison of ideal-actual analysis to classical and matrix sorting 56
Summary 67

V CONCLUSIONS AND RECOMMENDATIONS 68
Analysis of Specific Items 68
Items meeting criteria for difficulty level, point biserial correlation and distractor discriminating power 69
Items meeting criteria for difficulty level and point biserial correlation 69
Items meeting criterion for difficulty level 70
Items meeting criterion for point biserial correlation 70
Items meeting no criteria 71
Conclusions 72
Discussion and Implications 72
Difficulties in analysis procedure 73
Suggestions for Further Study 74
Summary 76

* * *

References 77
Bibliography 83
Appendix I Types of Examination Items 95
Appendix II Adapted Item Analysis Results 99

LIST OF TABLES

TABLE	DESCRIPTION	PAGE
I	Distribution of Tested Items	34
II	Classical Item Analysis Data - Group I	50
III	Significant Difference Values (z) Between Ideal and Actual Item Analysis Results	57
IV	Item by Item Comparison of Three Analysis Methods	61

LIST OF FIGURES

FIGURE	DESCRIPTION	PAGE
I	Histogram of Combined Group Results	38
II	Histogram of Group I Results	48
III	Matrix of Difficulty Levels for Graduating and Beginning Students	53
IV	Matrix Sorting of Item Difficulty Levels	55

I. Introduction

A school of nursing has a responsibility to its students, to itself and to the general public to ensure that only those nursing students who are safe, knowledgeable practitioners are allowed to write registration examinations as administered by the Canadian Nurses Association Testing Service (CNATS). To carry out this responsibility the school must evaluate its students in all aspects of nursing; that is cognitive, affective, and psychomotor . However, it is only the cognitive aspect of these three domains that the registration examinations ultimately test by means of multiple choice examinations. Since these registration examinations are administered on a nation wide basis, they must focus on knowledge which is common to all nurses in Canada; that is, the core of nursing knowledge.

A second responsibility of the school is to prepare its students to the level at which they are able to pass the registration examinations. It would seem that this end could in part be accomplished by giving students opportunities during their educational programs to answer multiple choice items which are similar to those on the registration examinations; and also to assist them to diagnose their strengths and weaknesses in relation to the core of nursing knowledge.

In order to carry out this dual responsibility of providing feedback to the students and assessing competence in nursing knowledge, some form of evaluation is necessary. Testing of the cognitive component can be achieved through the administration of multiple choice examinations in nursing. Through these the student gains experience in answering multiple choice items, the instructor and students are able

to diagnose areas requiring remediation, and the school administrators have concrete data on which to base decisions as to which students should be allowed to write the registration examinations administered by the CNATS.

Inherent in this position is the assumption that the multiple choice items on which the graduating students are tested are valid items. That is, the items exhibit the four types of validity; construct, content, concurrent and predictive. First, the items are able to differentiate between those students who have mastered the content essential to nursing and those who have not mastered this content; second, the items reflect the body of knowledge which is essential and particular to nursing; third, the student's performance on the items reflects the knowledge demonstrated by the student in the clinical situation; and fourth, the items are similar to those used on the CNATS registration examinations in nursing.

Content validity, that is the measure of the item's reflection of essential nursing content, can best be achieved by comparing items to instructional objectives, or by having practicing nurses evaluate the subject of the item in light of current work experience. As Cronbach (1971), Popham and Husek (1969), and Sax (1974) state, the established method for item formulation is to match items to the knowledge objectives of the course of study. However, as no formally documented core of nursing knowledge exists in Canada at present, there is no basis on which to establish the content validity of examinations used in preparation for registration examinations.

An item's utility in discriminating between competent and non-competent students is achieved by comparing the examinees'

performance on an examination to some type of standard. This standard might be an absolute one which is established by ensuring both content validity and clue-free construction of the items. In this case students achieving above a set cutting score or range of scores are termed competent, and those not achieving this standard are deemed non-competent. A problem exists in this method of standardization as classical test theory analyses methods are based on the assumption that the performance of subjects approaches a normal distribution, a distribution having a high degree of variance; whereas, by the time students are nearing the end of an instructional program which aims at competency, the subjects' performances more likely will be positively skewed, show little variance, and will not be normally distributed.

A much more frequently used measure of item quality is based on a relative standard. In this case examinees completing a test are compared to each other, and an interpretation of the results is based on this comparison. Each item in the test contributes to the comparison and is evaluated in terms of its contribution to the overall discrimination between competent and non-competent examinees. Statistical measures have been developed which aid in the evaluation of such items. For instance, a measure of the relative difficulty of an item is one criterion which can be used. An item which the majority of students answer correctly might be evaluated as an easy item, perhaps because it is poorly constructed, perhaps because it tests general knowledge rather than knowledge specific to nursing, or perhaps because it tests essential nursing knowledge with which all examinees are familiar. Similarily, correlations can be used to

determine the extent to which each item contributes to the entire test distinguishing between high and low ranking groups. An item might have a high correlation to the overall ranking of examinees, or it may have a small or negative correlation. Negatively correlated items indicate that there is some inherent fault; whereas small correlations may indicate a fault, or lack of knowledge by all examinees, or knowledge common to all examinees.

Another problem arising from the interpretation of the statistical analysis is the determination of the causal factor or factors leading to the students' performance on a particular item. Factors may be external to the item itself and in actuality be due to the persons being tested. If characteristics of the subjects are held constant, or built in to a particular group, interpretation of the above statistics becomes much more meaningful. For instance, if a group is constructed in such a way that it is comprised of known competent and non-competent subgroups, it is possible to establish an "ideal" performance for that particular group. This "ideal" performance is calculated using a hypothetical group in which competent students answer items correctly and non-competent students answer items incorrectly. Those items which act in the "ideal" manner can be identified, and those which do not act in this manner can be set aside for revision.

Statement of the Problem

This study is being conducted in order to establish a method for evaluating test items based on their utility in differentiating between examinees who possess competence in nursing, and those who do not possess this competence. More specifically, the individual

problems are to:

1. devise a methodology whereby test items are evaluated in relation to their utility in differentiating between a known group of graduating nursing students, and a known group of beginning nursing students in diploma programs in Alberta;
2. establish whether or not individual test items are able to differentiate between the known group of graduating nursing students and beginning nursing students;
3. and determine which specific analytical procedures provide information about the utility of individual test items in differentiating between the two groups.

Need for the Study

The University of Alberta Hospital School of Nursing is currently using its item bank to produce examinations at various points in its nursing education program. One set of examinations is administered prior to graduation in order to assist the students to prepare for writing the registration examinations. This set of tests covers content areas included in the registration examinations. Since, at present, there is no outline of what is considered to be core nursing knowledge within these content areas, selection of appropriate items is based on the judgement of individual nursing instructors.

The item analysis method used in this study could be used to

assist those producing the examinations to select those items which would indicate nursing competence, and in revising those items which do not meet the criterion of testing competence. The tool could also be used in a combined effort by all schools of nursing which prepare students for registration examinations. Each school could test items contained in its own examination item bank, and the ideal items then could be combined in a province-wide bank of test items which could be used to test knowledge prior to the CNATS registration examinations. It might also be feasible to adapt the methodology by changing the particular subgroups, to evaluate items for various other purposes, such as to distinguish between items appropriate to different levels of the same program.

Delimitations

The group of 128 test items used in this study were selected at random from the 1882 items contained in the University of Alberta Hospital School of Nursing test item bank. The items from this bank are currently used to test nursing students in a hospital based school of nursing in Alberta which prepares students for registration. These items were drawn at random from medical, surgical, pediatric, and obstetric sections of the bank. They also cover four types of multiple choice items, categorized along two factors; non-situational or situational, and multiple component or single answer.

The groups of nursing students involved in the study are from two schools in Edmonton, Alberta. Both schools offer basic diploma programs in nursing leading to eligibility to write registration examinations as administered by the CNATS. One of the schools offers

a three-year hospital based program; whereas the other offers a two-year community based program.

Limitations

The methodology as developed is only generalizable to multiple choice examinations in nursing which have a single correct answer. The study is further limited by the fact that mental health items are not included in the item bank as yet, and thus are not included in the test examination. Further limitations arise because the sample sizes of both items and students are small in relation to their respective populations; and the student subjects are drawn from a very narrow subset of the population. By nature of the establishment of a test group with known characteristics of competency and non-competency in nursing, the resultant distribution of scores is bimodal rather than normal as assumed in the statistical analysis.

Assumptions

In order to carry out this study, it is necessary to make the following assumptions.

1. It is assumed that the Canadian Nurses Association Testing Service registration examinations test areas of knowledge common to all basic nursing education programs in Canada.
2. It is assumed that the 72 subjects in Group I who are within one month of completing their basic education in Alberta are representative of all nursing students in Alberta who are about to write the CNATS

registration examinations in nursing.

3. It is assumed that the 93 subjects in Group II who are enrolled in the first month of a basic nursing education program in Alberta are representative of all students entering basic nursing education programs in Alberta.
4. Since the members of Group I have received nursing instruction, and since the members of Group II have received no instruction in nursing, it is assumed that the members of Group I are more competent in nursing than the members of Group II.

Questions to be Answered

In order to solve the problems as outlined, this study was directed towards answering the following questions.

1. Would a group of test items specific to nursing be able to differentiate between an known group of graduating nursing students, and a known group of beginning nursing students?
2. Could an item analysis technique be adapted from a classical test theory analysis technique in such a way as to compare ideal values for difficulty level, point biserial correlation, and distractor discriminating power with actual values for these factors for items tested on a known group of competent and non-competent students?

3. Would an item analysis technique which compares actual and ideal values for difficulty level, point biserial correlation, and distractor discriminating power provide more precise information for item revision than a classical item analysis method?
4. Would an item analysis method which compares actual and ideal values for difficulty level, point biserial correlation, and distractor discriminating power provide more precise information for item revision than a technique based on a matrix sorting of the difficulty levels for items tested on known groups of competent and non-competent students?

Definition of Terms

Terms and phrases used in this study are operationally defined in the following paragraphs.

1. Common Nursing Knowledge:

   Common nursing knowledge refers to that knowledge which is common to all basic nursing education programs, but which is separate from prenursing education knowledge, as shown by the ability of the graduating nursing students to correctly answer the item testing said knowledge, and the inability of the beginning students to answer the item correctly.

2. Graduating Nursing Students:

Graduating nursing students are those individuals who are engaged in their final year of studies at a diploma granting program in nursing, which leads to eligibility to write registration examinations as administered by the Canadian Nurses Association Testing Service.

3. Beginning Nursing Students:

Beginning nursing students are those individuals who are engaged in their first month of studies at a diploma granting program in nursing which leads to eligibility to write registration examinations as administered by the Canadian Nurses Association Testing Service.

4. Difficulty Level:

Difficulty level (p_i) refers to the proportion of students who correctly answer an item. It can vary from 0.000 to 1.000, and is computed by means of the formula $p_i = p/n$; where p is the number of subjects answering an item correctly, and n is the total number of subjects answering the item (Henrysson, 1971, p. 139).

5. Ideal Difficulty Level:

The ideal difficulty level for items is calculated from a hypothetical group of

competent and non-competent students. Allowing for a 10% chance of error, the percentage of competent students answering an item correctly was set at 90%; while the percentage of non-competent students answering an item correctly was equal to that of chance, that is 25% for a four distractor item, and 20% for a five distractor item.

6. Point Biserial Correlation:

Point biserial correlations measure the extent to which an item portrays the total test results by comparing the examinees' performance on an item with their performance on the examination as a whole. The range for the correlation is -1.000 to +1.000. Point biserial correlations are calculated using the formula:

$$r_{pbi} = \left[(\bar{X}p - \bar{X})/Sx\right]\sqrt{p/q}$$

where p is the proportion of correct answers to the item, q is the proportion of incorrect answers to the item, $\bar{X}p$ is the mean on the total test for those answering the item correctly, $\bar{X}$ is the mean on the total test for those answering the item, and Sx is the standard deviation of the test scores (Ferguson, 1976, p. 417).

7. Ideal Point Biserial Correlation:

The ideal point biserial correlation is based on a hypothetical group of competent and non-competent nursing students. It is calculated by using the above formula, and allowing for a 10% chance of error among the competent students; a 25% chance of correctly answering a four distractor item, and a 20% chance of correctly answering a five distractor item among the non-competent students.

8. Distractor Discriminating Power:

Distractor discriminating power is a measure of the ability of an item distractor to distinguish between those examinees who do well on the total test and those who do poorly on the total test. The range for the discriminating power is -1.000 to +1.000. Discriminating powers are calculated using the formula D.P. = H - L; where H is the proportion of high scoring examinees who chose the distractor, and L is the proportion of low scoring examinees who chose the distractor (Henrysson, 1971, p. 145).

9. Ideal Distractor Discriminating Power:

The ideal distractor discriminating power is based on a hypothetical group of competent and

non-competent students. The calculation allows for a 10% chance of error among the competent students, and a chance proportion of correctness among the non-competent students; that is, 25% for a four distractor item and 20% for a five distractor item.

Summary

As the University of Alberta Hospital School of Nursing test item bank is being used, it is becoming evident that the items need to be evaluated as to their usefulness in performing certain specific tasks. It is the purpose of this study to develop a method of evaluating items as to their utility in distinguishing between those students who are competent and those who are not competent in the field of nursing. Theories of testing and item analysis on which the adapted methodology is based are outlined in the next chapter. Chapter 3, in turn, delineates the mechanics related to the development of the instrument and evaluative methodology; while results of the implementation of the instrument, and implications arising from this are reported in chapters four and five.

II. Review of Related Literature

Education, as defined by Krathwohl and Payne (1971) is "a process of changing student behavior to achieve certain specified goals" (p. 18). Testing thus becomes a means of determining whether or not these goals have been attained, or in other terms, a means of decision making. Cronbach and Gleser (1965), Ebel (1965) and Glaser (1963) reiterate this functional definition of testing with the added stipulation that the quality of the decision is directly related to the type of measurement used in the test itself, and the extent of its validity for the type of decision needed. That is, the quality of the decision is directly related to the test type and structure.

Therefore, in this review of the literature relevant to the study of test theory, an initial step is to explore the uses of tests in relation to the types of decisions required. This is followed by a review of test types, their formulation, score interpretation, and item analysis.

Uses and Structure

Manning (1965) broadly categorizes three types of tests; those used for decisions related to selection and distribution, those related to diagnosis and prescription, and those for evaluation (p. 14). Authors such as Dressel (1964), Ebel (1965), Klein (1970), Lindquist (1951), Sax (1974) and Thorndike (1971) outline various uses of tests in more specific terms, all of which fit into these three categories.

Selection and distribution decisions may be made to certify competence in a training or professional field, select students for

admission to an institute of higher education, award scholarships, or assign students to particular types of curricula. Diagnosis or prescription tests are used to motivate and direct learning to least known areas, determine remedial learning needs, give feedback to students, or in themselves teach. Promotion and grading practices fall into the category of evaluation.

The variation in test use necessitates differences in test structure, so that the specific test best serves the intended purpose. In general, two basic structures in testing have emerged; the first being a norm referenced, and the second being a criterion referenced form of test. The majority of the literature within the past two decades has focused on the uses, abuses, and difficulties arising from the implementation and interpretation of criterion referenced measurement. Yet, as norm referenced testing is more commonly understood by the classroom teacher, most authors use it as a foundation for discussion. This particular format will be adhered to in this review.

Norm and criterion referenced testing. Norm and criterion referenced tests differ basically in the standard to which the examinee's achievement is compared. There is general agreement among authorities in the testing field that a norm referenced measure is used to compare an individual's performance in relation to that of others on the same test; and a criterion referenced measure is used to identify the individual's performance in relation to a pre-established standard of performance (Esler and Dzubian, 1974; Glaser, 1963; Hambleton and Novick, 1973; Hunt and Randhawa, 1976; Popham and Husek, 1969; Thorndike, 1971). By using this distinction between the two test types,

it is possible to reclassify the uses of tests as reported by Manning. Criterion tests are used for diagnostic purposes, evaluation, to ascertain the need for remediation, and in some instances, to allow certification of minimum competence in a field (eg. driver training). Norm reference tests are traditionally used to grade and promote students, assess entrance qualifications to institutes of higher learning where quotas of student positions are in effect, and award scholarships.

Although an overlap exists in the use of both types as measures for promotion and assessment of scholastic achievement, Shoemaker (1975) summarizes the recent trend from the traditional normative form to the criterion referenced mode:

> There is the furor over criterion-referenced achievement testing and the concomitant dissatisfaction with standardized achievement tests. There is the public demand for concrete evidence that the monies being spent on education are being spent well. And there is the necessity for a precise assessment of student achievement growing out of the desire by goverment, both state and local, to allocate resources in accordance with educational need. (p. 128)

Yet he and Dzubian and Vicker (1973) point out that much of the reluctance to change is due to the classroom teacher's security in traditional methods, and the lack of easily interpreted evaluative measures for the criterion referenced tests.

Formulation of Tests

As early as 1931, Ralph Tyler proposed a ten step process for the formulation of tests. The steps are as follows:

1. formulation of course objectives
2. definition of objectives in terms of student behaviors
3. collection of situations in which students will reveal the presence or absence of each objective
4. presentation of situations to students
5. evaluation of student reactions in light of each objective
6. determination of the objectivity of evaluation
7. improvement of the objectivity as required
8. determination of reliability
9. improvement of reliability as required
10. development of more practicable measurements as required

These ten steps have been adhered to through the years since they were first published, by constructers of both norm and criterion referenced tests (Ebel, 1971; Glaser, 1971; Thorndike, 1971; Vaughn, 1951). The difference between the use for the two tests tends to centre in the level of specificity of interpretation of the first three steps, and in the fifth step.

Relationship to course objectives. Hively (1974) in his text on domain referenced testing describes this phenomenon of multivariate interpretation very well. He describes a task of improvement in the skill of shooting flying targets. In the domain referenced mode, (criterion mode), the test situation is clearly outlined to the subject prior to the examination, and direct feedback is given as to the specific areas needing improvement. No comparison is made to

other shooters testing their skills in the same situation. In the norm referenced mode a much more generalized idea of the test situation is given, and feedback is directed to the examinee's ability in comparison to other shooters, but not to specific skill areas needing improvement.

At this point, it seems appropriate to clarify some of the terms synonymous with criterion referenced testing. In various texts and articles, criterion referenced tests are alternately referred to as competency based, tests for mastery, and domain referenced (Bloom, 1973; Hively, 1974; Millman, 1974). Popham (1975) seeks to clarify this potpourri of terms by his functional definition of a criterion referenced test as one which "is used to ascertain an individual's status with respect to a well-defined behavior domain" (p. 130). From hereon, the term criterion referenced test will be used to refer to any of competency, mastery, or domain referenced tests.

In actuality, the term domain referenced test arose from difficulties encountered by early theorists in delineating the entire population of test items which would be measures of a particular behavioral objective. The examinee's performance on the entire population of items is the best measure of an objective. However, this population of items is often too large to provide a feasible test situation. Therefore, an estimate of the examinee's ability with respect to the objective is made from his ability on a random sample of test items from a clearly defined domain of items which measure the objective (Hambleton, Swaminathan, Algina, Coulson, 1978, p. 3). In many circumstances, test items are referenced to objectives, but as no domain of behaviors is specified, the test, by definition, is not

criterion referenced. This particular circumstance may arise in the formulation of norm referenced tests, and follows the first two steps as identified by Tyler. In this particular situation Hambleton et al (1978) state:

> that when items in a norm-referenced test can be matched to objectives, criterion-referenced interpretations of the score are possible, although they are quite limited in generalizability. A criterion-referenced test constructed by procedures especially designed to facilitate criterion-referenced measurement can and sometimes is used to make norm-referenced measurements. (p. 3)

The authors proceed to state that these are not the most satisfactory situations however, and that decisions made under these circumstances are somewhat tenable.

Item reliability. A second distinction in the formulation of a criterion referenced versus a norm referenced test, centers in the interpretation of steps six and eight of Tyler's format, that is, in the determination of the objectivity and reliability of a test. If the purpose, as in norm referenced testing, is to compare examinees, test items must be formulated to maximize minute differences in knowledge and thus spread examinees on a linear continuum of achievement. In criterion referenced tests the minute differences between examinees lose their meaningfulness. Therefore variance based correlations which are used in the analysis of reliability of norm referenced tests are not applicable in criterion referenced tests (Dzubian and Vickery, 1973; Hambleton and Novick, 1973; Hunt and Randhawa, 1973; Lord and Novick, 1968; Popham and Husek, 1969).

A more extensive review of this area is included in the section dealing with the difficulties arising from the interpretation of test scores. Yet it is an important means of evaluating the quality of items being used to make up a test, and thus in the selection of those items.

Interpretation of Test Scores

Popham and Husek (1969) succinctly examine the reporting and interpretation of test results in relation to the type of decision-making information gained from the test. Since norm referenced tests are used for group comparison decisions, results are published in comparative terms such as percentiles, ranking, stanines, or some other form of standard score. The drawback to this system is that little is known about the absolute ability of the individual examinee. In an attempt to increase the meaningfulness of the comparative score, range, mean and standard deviation are often reported as well (Ferguson, 1976).

Interpretation and reporting of criterion referenced test scores are made with reference to some specific mastery score (or a set of these scores). The individual's score is reported in terms of achievement or non-achievement of some pre-determined set of mastery. In the ideal situation, mastery would be the state of answering all items within a specified domain correctly. To allow for some error factor, the mastery level is often set slightly below 100%. The use of several cutting scores allows for allocation of non-mastery examinees to different levels of remediation treatments (Popham and Husek, 1969). Again, this particular method of score interpretation

is incomplete in many cases, in which an estimate of comparative expertise is desired. In this case group scores may also be reported by a term such as 92-90, meaning that 92% of the group achieved 90% or better on the examination (Popham and Husek, 1969, p. 8).

Ebel, (1962, p. 15) emphasizes that in criterion score interpretation a separate score must be reported for each domain tested. If a compilation is reported, the examinee is not able to discriminate areas of non-mastery from those of mastery. Hunt and Randhawa (1976) report four ways of expressing criterion referenced scores.

1. The number or percentage correct on a given objective or set of items that encompass a few highly related objectives.
2. "Master" of a given objective or set of items where "mastery" is defined in terms of certain levels of performance such as 90% correct.
3. The time it takes (class hours or days) for an individual to achieve a given performance level.
4. The probability that the student is ready to begin the next level of instruction. (p. 11)

In his writings from 1962 to 1971, Ebel reports that no matter which form of testing is used, the score, to be meaningful must relate both to the content of the course, and to the examinee group's performance.

True score. Aside from the meaningfulness of a score, another difficulty in the interpretation of both norm and criterion referenced

test scores lies in the use of the examinee's raw or observed score as an estimate of his true score. True score is defined as "the average of the scores that the examinee would make on all possible parallel tests if he did not change during the testing process" (Lord, 1967, p. 41). In more practical terms, the true score is the reflection of the examinee's actual knowledge and ability, uncompromised by any external factors. This concept applies equally as well to norm as to criterion referenced scores. The examinee's raw score is thus made up of two components, his true score and his error score. It is the error score that is due to factors external to his actual ability and knowledge (Jones, 1971). His true score can be expressed as a linear equation $T = X - E$, in which X, the observed score, and E, the error score are both variable, but in which T, the true score is constant for the examinee (Gulliksen, 1950; Lord, 1967; Lord and Novick, 1968).

Lord and Novick (1968) report that there is some disagreement as expressed by Loevinger (1957) and Thorndike (1964) on the importance of the concept of a true score, as it can only be estimated and not specifically determined. Yet Gulliksen (1950) uses a reduction in the variance of error scores as a cornerstone for improving a test. "Much of the effort in test construction, test revision, test analysis, and the precautions of test administration are for the purpose of decreasing the value of Se (error of measurement)" (p. 11). General causes of and ways to reduce the error of measurement are reported in studies of three basic types of errors. The first type refers to those errors within the examinee, such as fatigue, lack of motivation, test wiseness, or anxiety; the second are within the test itself, such as

ambiguity, poorly worded directions, and poorly constructed items; and the third type of factors are those found with the scoring conditions, such as carelessness, a lack of set standards, or computational errors (Alker, 1969; Diamond, 1972; Hopkins, 1972; Sax, 1974; Strang, 1977; Wright, 1967).

Further in the attempt to reduce the error measurement as much as possible is the study of the measurement instrument itself. These studies focus on two main concepts, first that the instrument is accurate in measuring what is intended to be measured (validity); and second, that the scores will be reproduced exactly on remeasurement (reliability).

Validity. Although the concept of validity seems to be axiomatic to test theory, there are many varied interpretations of the term. In 1961, Ebel identified six unique definitions of the term (p. 220). Based on the differences between the definitions, he goes on to suggest that validity as a concept is not important, and that the test interpreter would do well to focus on the meaningfulness of test scores as a measure of test quality. The American Psychological Association, in an attempt to standardize some of the work on validity published a document in 1954 which outlines four types of validity (pp. 13-28). Content validity indicates how faithfully the objectives and emphasis of a course of instruction are represented by the test. Predictive validity indicates the test ability to indicate future traits. The validation is based on the correlation of test scores with evidence of the trait gathered at a later date. Concurrent validity is similar to predictive validity except that the correlation is made between the test and evidence of the trait gathered at the same time. These two

latter types of validity are sometimes grouped under the term of criterion related validity (Cronbach, 1971, p. 444; Sax, 1974, p. 209). Construct validity refers to the quality of a test in inferring what psychological traits or constructs a test measures, such as the use of the CNATS registration examinations to measure the examinee's ability to implement safe nursing care.

Analysis of validity is done through correlational studies, and as such is dependent on some variance within the abilities of the subjects being tested. Therefore, measurements of validity are more meaningful on norm referenced tests than on criterion referenced tests (Hambleton and Novick, 1973; Popham and Husek, 1969). Popham and Husek do not see this as a shortcoming of criterion referenced test scores, though, as the necessary foundation of criterion referenced tests is that they adequately represent the domain of behaviors under examination; that is, that they have a high degree of content validity (1969, p. 6). The American Psychological Association indicates a drawback with the estimation of content validity in that "in most classes of situations measured by tests, quantitative evidence of content validity is not feasible" (1966, p. 3). Cronbach (1971) and Sax (1974) suggest that content validity can be quantified by having two independent groups of item writers construct tests to meet a given format, set of objectives, and number of items per objective; administering both forms of the test to a group of examinees, and correlating the results.

In all, the experts agree that validity is axiomatic to a quality test, but that the presence and estimation of the validity are not easily attained.

Reliability. The second characteristic of a test which is applicable to the accuracy of the raw score as an estimate of the true score is that of the reliability or consistency of scores from one set of measures to a parallel set. Stanley (1971) suggests that the causes of error scores as mentioned earlier; that is, those inherent in the individual, those in the test, and those in the scoring conditions, lead to unreliable scores. He emphasizes the first two categories, however.

Reliability measures are estimated through correlational studies between at least two scores per subject over a group of subjects. Gulliksen (1950) and Stanley (1971) outline three methods of obtaining this correlational data. They advocate the use of parallel test forms, a test-retest situation, or using split halves of the same test for correlations. The various formulae outlined in basic statistics texts, such as Ferguson (1976), Gulliksen (1950), and Lord and Novick (1968), can be used on the scores of both norm and criterion referenced tests. However, the interpretation of those results for criterion referenced scores is somewhat suspect because the coefficient of correlation is related to the variance of scores among different examinees (Stanley, 1971).

Noting this difficulty, researchers have devised various other means of determining the reliability of criterion referenced scores. Livingston (1972) suggests a study based on group variance from the cutting or mastery score of the test. Hambleton and Novick (1973) query the usefulness of this measure though, as the purpose of the test is not to evaluate the extent of variance of an examinee from the cutting

score, but rather to indicate mastery or non-mastery. Brennan and Kane (1977) follow a similar method in determining an index of dependability for mastery tests. Subkoviak (1978) compares four methods those of Swaminathan, Hambleton, and Novick; Huynh; Marshall and Haertel; and Subkoviak for determining the consistency of assigning mastery and non-mastery states to examinees. Popham and Husek (1969) suggest that a forced variance in test scores be introduced by doing correlations between performance on equivalent pre and post instructional test forms. An alteration to this method has been used by Haladyna (1974) and by Kosecoff and Fink (1976). They correlate the scores of known mastery and non-mastery students on the same test form.

As with Ebel's (1962) caution of reporting a separate score for each domain, Popham and Husek (1969) note that care must be taken to estimate the reliability for each domain within a test.

Item Analysis

As stated previously, the examinee's error score is related to the test validity and reliability. Factors contributing to low validity and reliability can be inherent in the examinee, the examination administration conditions, or the examination itself. Much has been done in test theory through the use of item analysis to improve the test items themselves. Gulliksen (1950), Hubbard and Clemans (1961), Lippey (1974), Popham and Husek (1969), and Sax (1974) state that traditionally item analysis techniques have been used to identify items which do not properly discriminate between examinees on a particular test. "Proper" discrimination is identifiable in

items which are neither too easy, too hard, nor ambiguous. Although different statistical methods exist for analyzing items, they can generally be classified into two categories. Historically, when computers were not readily available, classical test theory techniques were used (Henrysson, 1971). As more advanced calculations and computers became more available, the latent trait models were developed (Lord and Novick, 1968; Henrysson, 1971). Although item analysis techniques indicate items in which there may be faults, they fall short of indicating the particular fault (Hubbard and Clemans, 1961).

Classical test theory. Classical test theory stresses three measurements which are performed on each item. The difficulty level (p), or proportion of examinees correctly answering the item is the first of these measures. Henrysson (1971) suggests that a more satisfactory measure of item difficulty is based on the assumption that the examinees' ability to answer a question is not dichotomous, but rather normally distributed. In light of this assumption, Henrysson suggests that a z-score calculated from p and a set of normal curves be used for the difficulty level.

A second measure is that of item discrimination power, which is usually calculated through an adaptation of the Pearson Product Moment Correlation. Most common in usage are the biserial and point biserial correlations (Gulliksen, 1950; Henrysson, 1971; Lord and Novick, 1968). The relative worth of these two discrimination indices, as well as others such as tetrachoric and phi coefficients, has been compared by Oosterhof (1976), who finds little to indicate

the use of any one index over another.

In order to increase the meaningfulness of the item difficulty and item discrimination indices, which are to some extent interdependent, Hofmann (1975) has developed an item analysis index. This index is a function of both the item difficulty and item discrimination indices and can be used to compare the actual efficiency of an item with the maximum efficiency possible for that item (p. 626).

Once the general quality of an item is determined, the quality of each distractor can be assessed by checking the mean scores of the groups of examinees who chose the distractor. The examinees' mean score on the erroneous answers should be lower than the mean of the group choosing the keyed answer (Henrysson, 1971). A second option is that of comparing the proportions of high scoring examinees to that of low scoring examinees for each distractor (Gulliksen, 1950). Feldt (1961) suggests that the high and low group sizes should be 25 to 27% of the entire population in order to provide powerful statistical data.

In relation to criterion referenced tests, Popham and Husek (1969) suggest that care must be taken in the interpretation of these item analysis techniques. Erroneous interpretations may arise if an item is judged as non-discriminating because it is too easy, when in fact it is reflecting the information that all examinees have the knowledge to answer that question. This, in fact, is the problem outlined in this paper, as items show little discrimination between students who are nearing completion of a nursing program, and who have relatively the same knowledge base. Popham (1974) also indicates that care must be taken when using classical item analysis

techniques to interpret a norm referenced test as well.

> In many cases the very items which deal with the concepts or skills considered most important to a field, will, over time, be systematically eliminated from a norm-referenced achievement test. (p. 614)

Latent trait model. The latent trait models are discussed by Lord and Novick (1968) and Baker (1977). They assume that performance on test items is dependent upon some underlying trait. The item characteristic curve is "the function of relating probability of success on the item to the examinee's position on the underlying trait" (Henrysson, 1971, p. 146). This method of item analysis has the advantage of showing the discriminating ability of an item in a graphic form which clearly indicates trends in discriminating ability, and changes in those trends. The disadvantage of this type of item analysis is the complicated and time consuming computational work which necessitates the use of a computer. This leaves the classical method as the most advantageous one for the classroom teacher.

Item banking. Henrysson (1971) and Sax (1974) both advocate the use of a central file or bank of test items and their respective analyses. Through the use of such a bank, item selection for particular tests can be aided by a review of the analyses from the item's use on former tests. A word of caution is given by Lippey (1974) in the use of stored items. He states:

> Since the obtained values of item statistics are situationally dependent, their use in the item improvement cycle requires that the context in which the statistics were calculated be known. (p. 157)

Summary

In the past two decades the development of test theory has tended to be somewhat cyclical in nature. The advent of individualized instruction has brought about the necessity of an individually based evaluation system, that is criterion referenced testing. This, in turn, has led to the development of new theories in an attempt to analyze the value of criterion referenced tests and items. The usefulness of the classical test model has been questioned at this time due to its reliance on group performance and the variability of the members within the groups. The difficulties arising in the analysis of criterion referenced testing have led to discussions of the relative merits of norm and criterion referenced testing, and of which decisions are congruent with which type of test.

Superimposed on these issues has been the increasing availability of computers with which to analyze test data. This, in itself, has led to new developments in item analysis techniques, and to refinements in the classical theory techniques.

At present, the entire field of test theory seems to be going through a period of dynamic change, which lends itself to the development of new techniques and/or variations in existing ones in the furtherence of test theory knowledge.

III. Methodology and Instrumentation

Questions have been raised at the University of Alberta Hospital School of Nursing as to the quality of the items used on examinations. Items are analyzed by the classical test theory method as they are used, yet the technique has had many of the pitfalls outlined by Alker (1969), Diamond (1972), Hopkins (1972), Popham (1974), Sax (1974), and Wright (1967). The test item may be shown statistically to have faults, but the origin of the fault or faults to the item or to the subject group is not known. For instance, the group of examinees may have the same basic knowledge, and thus little variance amongst them. Items analyzed in this context may appear to have internal faults, as seen in high difficulty levels, low correlations, and low distractor discriminating ability, when the factor influencing the analysis is the external homogeneity of the group itself. Therefore decisions regarding the quality of the UAH School of Nursing test items have been difficult to make due to a lack of a sound decision theory base.

Popham and Husek (1969), Haladyna (1974), Kosecoff and Fink (1976) suggest that the use of known differing subgroups of examinees increases the reliability of item analysis statistics by controlling for the homogeneity of the examinee group. A second benefit to the use of a known group of examinees for testing examination items is the fact that the ideal manner in which an item should perform could be precalculated and used for comparison sake.

These two factors have influenced the development of the item evaluation methodology outlined in this chapter.

Development of Item Bank

The UAH School of Nursing computerized item bank was developed in 1975 and 1976. The faculty members perceived a need for a centralized bank of multiple choice items which would be available to instructors and which would facilitate examination production. An advantage to the computerized bank was seen as it's being a secure, centralized storage area for items, which would allow for examination production in a short time with little consumption of secretarial time. As editorial and typing times are decreased, it was foreseen that instructors could focus on improving item quality. Item analysis data which would aid in this process would be readily available with the item in the bank. A third advantage to this particular system was the availability of diagnostic feedback for both students and instructors.

As a starting point to the development of the bank, all multiple choice items in the school were collected and categorized into six main categories; medical, mental health, obstetric, pediatric, and surgical nursing; and nursing arts and sciences. Each of these was to have its own file developed. The original pilot file was completed in obstetrical nursing, with medical, surgical, and pediatric nursing files following. The remaining two files have not as yet been developed. At present, there are 499 items in the medical nursing file, 286 in the surgical nursing file, 516 in the pediatric, and 581 in the obstetric nursing file.

The particular item banking program in use allows a variety of information to be stored. Some files contain the items themselves, some actual examinations, some the data from student answer sheets,

and some the analysis of the items used on a test. Subprograms allow for the production and analysis of an examination. Hard copy printouts of each item with information as to its year of entry, number of times used, keyed distractor, and current analyses are also available through this program.

The majority of the items contained in the item bank follow one of two construction formats. These are single answer or multiple component. The single answer item is similar to the type "a" item outlined by Hubbard and Clemans (1961). It differs in that the type "a" item is standardized to five distractors, and the items in the UAH School of Nursing item bank may have three, four, or five distractors. With this type of item, the examinee selects the one distractor which best answers or completes the stem.

The multiple component item follows the same basic format of Hubbard and Clemans' (1961) type "k" item. These authors have refined their items so that the type "k" item has a standard set of secondary distractors. However, the multiple component items in the UAH School of Nursing item bank are not so standardized. In answering an item of this type, the examinee chooses a number of primary distractors which best complete the stem, and then chooses a secondary distractor which gives the corresponding combination of primary distractors. Examples of single answer and multiple component items are presented in Appendix I.

In order to create a more realistic nursing situation, some of the items in the bank refer to case histories. These situations present the examinee with more information about a particular patient, and are followed by a number of items for which pieces of the given information

are needed. These items in the bank are termed situational items Examples of them are also presented in Appendix I.

Development of Instrument

As the bank contains a total of 1882 items, it was not feasible to use the developed methodology to test all items at this particular time. Therefore, the test examination was compiled in such a manner as to have equal representation from each file, and from each of the main item formats within the bank. Eight of each of the four formats were selected at random from each of the four item banks, resulting in an overall total of 128 items. Table I shows the distribution of the item types and files.

Table I

Distribution of Tested Items

Item Format / Item File	Situational		Non-situational		Total Items
	Single Answer	Multiple Component	Single Answer	Multiple Component	
Medical	8	8	8	8	32
Pediatric	8	8	8	8	32
Surgical	8	8	8	8	32
Obstetric	8	8	8	8	32
Total Items	32	32	32	32	128

The number of distractors per item was not controlled in the examination. The resultant test had 90 four distractor and 38 five distractor items. Once the items had been selected, they were randomly ordered on the examination. An exception to this ordering

occurred in cases in which two or more items referred to one situation. In this instance those particular items were grouped together. The examination paper was produced by computer from the item bank.

Administration of the Instrument

In May, 1978, the examination was administered to 52 nursing students at the Royal Alexandra Hospital in Edmonton. These students were within one month of completion of their nursing program, and were eligible to write registration examinations in nursing. In June of the same year, the examination was administered to 20 students at the Grant MacEwan Community College Nursing Program in Edmonton. As with the former group, these students were within one month of completion of their program, and were eligible to write registration examinations. The main difference between the two groups was that the first section of students were instructed in a three year hospital based program, and the second section in a two year community college program. The above students were classified as Group I for the study.

Students who wrote the examination did so on a voluntary basis. No time limit was fixed for the examination, and the examinees were requested to attempt all items. In order to maintain the confidentiality of the students, they were requested to identify their answer sheets by their school name, and graduating year only. Identification numbers were affixed to the answer sheets after the test administration.

Under the same conditions as above, the examination was

administered in September 1978 to 98 beginning students at the Royal Alexandra Hospital School of Nursing. These students were within one month of the start of their nursing education program and comprised Group II in the study.

Students from the University of Alberta Hospital School of Nursing were excluded from the study as they had been exposed to some of the test items on previous examinations within their program.

Examination Analysis

All test papers were optically scored by the IBM 1230 Optical Scorer at the University of Alberta. A total of 19 of the 170 answer sheets were rejected from the study, as the subjects had not completed the examination, or had filled in more than one answer per item. This left Group I with a total of 62 subjects, 43 from the Royal Alexandra Hospital, and 19 from the Grant MacEwan Community College. Group II was left with 89 subjects.

Item analysis. Item analysis data was computed by the TEST04 program of the Division of Educational Research Services, University of Alberta. Separate analyses were carried out on Groups I and II, and on the combined groups. This particular program provided information on the test as a whole; including the mean, variance, standard deviation, Kuder Richardson reliability coefficient, and a histogram of the examinees' results. Difficulty levels, point biserial correlations, item reliability indices, biserial correlations, and discriminating powers for distractors were provided for the individual items.

Comparison of means. A t-test was conducted comparing the means of Group I and Group II using the formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

where $\bar{X}_1$ = mean of Group I, $\bar{X}_2$ = mean of Group II, and $s_{\bar{x}_1 - \bar{x}_2}$ = the standard error of difference between the two means (Ferguson, 1976, p. 169).

Adapted Item Evaluation Methodology

The TEST04 program needed little adaptation to compare Group I and Group II. Both the difficulty level and the point biserial correlations were computed on data from the combined groups in their entirety. However, only the top and bottom 27 percents of the entire group were used in the TEST04 program for the calculation of the distractor discriminating power. This percentage has been demonstrated by Kelley (1939) and Feldt (1961) to produce the optimum comparison for a normally distributed subject population. As the distribution of scores within a combined group of known competent and known non-competent students was more likely to be bimodal and the 27 percent group captured neither mode, an adjustment was made to the program so that the high comparison group would include the majority of Group I, and the low comparison group would include the majority of Group II. This resulted in the high and low groups each containing 45% of the entire group. The information needed to make this adjustment was obtained from the histogram of the combined group results, as shown in Figure I. The cutting scores between the low, mid, and high groups

Figure I - Histogram of Combined Group Results
(Note: Group I = ·, Group II = *)

were 48.5 and 53.5 respectively.

Following the adaptation of the TEST04 program, item analysis was completed on the combined groups. The results of item difficulty levels, point biserial correlations, and distractor discriminating powers were obtained and further used in calculations comparing them to the ideals expected for items discriminating between competent and non-competent examinees.

Computation of Item Analysis Values for Ideal Items

Based on the work of Haladyna (1974) and of Kosecoff and Fink (1976) in using known mastery and non-mastery subjects to analyze test items and tests, a similar procedure was employed in this study. The question at hand was to evaluate the test items using a known group of students who were competent in nursing, and a known group of students who were not competent in nursing. In order to evaluate the actual test items, calculations were made as to the manner in which ideal items would perform in the given group. Competence was defined as having 90 percent of the items correct. This figure allowed the competent students some chance of error. Non-competence was defined as having a chance percentage of correctly answered items. In order to do computations of chance error, the examination was divided into its two components; that is, the component of 90 four distractor items and the component of 38 five distractor items. This was necessary as the non-competent group had a 25 percent chance of answering a four distractor item correctly, but only a 20 percent chance of answering a five distractor item correctly. Ideal item

difficulty levels, point biserial correlations, and distractor discriminating powers were computed for both four and five distractor items.

Four distractor items. Based on the three givens of a total of 90 four distractor items, a 90 percent chance of the 62 members of Group I answering any item correctly, and a 25 percent chance of the 89 members of Group II answering any item correctly, the following computations were made.

The difficulty level for an ideal item was computed as follows:

$$p_i = p/N = .517$$

where p = the number of subjects answering the item correctly

= 78.05

and N = the total number of subjects answering the item

= 151

The ideal point biserial correlation required more extensive calculations for the formula:

$$r_{pbi} = \frac{\bar{X}_p - \bar{X}}{s_x} \sqrt{p/q}$$

where $\bar{X}_p$ = ideal mean of those answering the item correctly

= 64.32

$\bar{X}$ = sum of the ideal means of Group I and Group II

= 46.52

S_x = standard deviation of an ideal examination

= 28.78

p = difficulty level = .517

and q = 1-p = .483

Therefore:

$$r_{pbi} = \frac{64.32 - 46.52}{28.78} \sqrt{.517 - .483}$$

$$= .640$$

The entire group of 151 subjects was used to calculate the distractor discriminating power for the keyed answer of an ideal item because an ideal item would have divided the combined group distinctly into known subgroups. The distractor discriminating power was:

$$H - L = .65$$

where H = proportion of the high group choosing the distractor

= .90

and L = proportion of the low group choosing the distractor

= .25

Five distractor items. Similar computations were made for the 38 five distractor items in the examination. The resultant ideal difficulty level was p_i = .487; the point biserial correlation was .688 and the ideal distractor discriminating power for the keyed distractor was .70.

Comparative analysis. From the adapted item analysis results a number of separate computations were made. The significant difference between the difficulty levels for the ideal and actual items were calculated by means of the formula for significant difference of two proportions (Ferguson, 1976, p. 174).

$$z = \frac{p_1 - p_2}{\sqrt{pq\,(1/N_1 + 1/N_2)}}$$

where p_1 = difficulty level for an ideal item

p_2 = difficulty level for an actual item

$$p = \frac{\text{actual plus ideal number answering the item correctly}}{N_1 + N_2}$$

$q = 1 - p$

and $N_1 = N_2$ = combined group size = 151

The same formula was used to determine the significant difference between the ideal distractor discriminating power and actual distractor discriminating power for each keyed answer. In this case:

p_1 = ideal discriminating power

p_2 = actual discriminating power

p = difference between numbers in actual high and low groups answering the item correctly plus the difference between the numbers in the ideal group answering the item correctly, divided by $(N_1 + N_2)$

$q = 1 - p$

N_1 = combined group size = 151

N_2 = high plus low ranking students in actual group = 134

Significant differences between the point biserial correlations for ideal and actual items were tested using the formula for the difference between correlation coefficients (Ferguson, 1976, p. 184).

$$z = \frac{z_{r1} - z_{r2}}{\sqrt{1/(N_1 - 3) + 1/(N_2 - 3)}}$$

where $z_{r1} = z_r$ for ideal correlation

$z_{r2} = z_r$ for actual correlation

$z_r = \frac{1}{2} \log_e (1 + r) - \frac{1}{2} \log_e (1 - r)$

$N_1 = N_2$ = combined group size = 151

From these calculations, assessments were made as to the item's discrimination between a known group of competent and non-competent nursing students.

Comparative Assessment of Individual Items

A comparison of three item analysis results was carried out for each item in order to determine which type of analysis provided the most pertinent information about the item. The first assessment was the classical item analysis of Group I. The analyses of Groups I and II were then used to do a crude sorting of items by difficulty level in order to determine which items discriminated between the competent and non-competent students. The third assessment was of the information provided in the adapted methodology which compared the adapted item analysis of the combined groups with the analysis expected for an ideal item.

Summary

In order to meet the need for more useful information concerning the quality of test items used at the University of Alberta Hospital School of Nursing, an adaptation to the classical test theory was employed. The new method, as suggested by Popham (1969), involved the testing of items on a group of subjects with known characteristics as to their competence or incompetence in nursing. After analysis of the items was obtained using classical test theory methodology, a

comparison of the actual item analysis data to ideal data as computed for the subject group was obtained to provide more substantial information.

The steps involved in the adaptation of the item evaluation methodology were the:

A. Compilation and Production of Examination

1. compilation of test items for examination
2. production of examination

B. Administration of Examination

1. administration of the examination to a known group of competent students
2. administration of the examination to a known group of non-competent students

C. Analysis of Items

1. classical analysis of competent group's tests
2. classical analysis of non-competent group's tests
3. classical analysis of combined group's tests
4. comparison of means of competent and non-competent groups
5. analysis of histogram of combined groups for appropriate cutting points for distractor discriminating power analysis
6. adaptation of classical analysis (TEST04) program
7. analysis of items using altered cutting points
8. calculation of ideal item analysis data
9. comparison of ideal and actual item analysis data

D. Evaluation of Items

1. assessment of individual items from analysis data of Group I
2. sorting of data from Group I and Group II to provide difficulty level matrix
3. assessment of individual items using data from significant difference analyses

These steps were initiated for a number of examination items in nursing, and the results were compared to classical item analysis information for the item. The information related to compilation, production, administration and analysis has been presented in this chapter. The following chapter deals with the last step, that of the assessment of the individual items, and their ability to differentiate between competent and non-competent nursing students.

IV. Analysis of Results

The previous chapter outlined different methods of collecting item analysis data. The data, though, lacked usefulness in that it had not been interpreted in light of the decision required; that is, the ability of an item to differentiate students who were competent in nursing from those who were not competent in the field. Initially, item analysis data, collected on the administration of the test examination to graduate students alone, was interpreted in light of the known characteristics of the subject group; that is, the assumed competence of this group of students. Second, the difficulty levels for graduating students and beginning students were compared for each item. The items were sorted into a nine cell matrix according to their respective difficulty levels. The matrix was used to determine which items would meet the criteria of differentiating between graduating and beginning students in such a way that the graduates found the item easy, whereas the beginning students found the item difficult. The third method used to evaluate the items compared the item's actual performance to a calculated ideal performance, given that the group of students on whom the items were tested consisted of known subgroups of competent and non-competent subjects.

Test Analysis Data Results

Prior to examination of the individual items by the three item analysis methodologies, the test data was examined to ensure that there was a significant difference between the results of the graduating and beginning students.

On the 128 item test, the graduating students' raw scores ranged from 34 to 87 with a mean of 71.27; while the beginning students' scores ranged from 24 to 59 with a mean of 45.31. The amount of overlap was negligible, with all but two of the graduating students obtaining more than 56, and all but one of the beginning students achieving less than 56. There was a significant difference between the two means with $t = 20.354$ ($\alpha \leq .001$).

The amount of variance reported for the graduating students was 77.46, while that for the beginning students was 33.59. On combining the groups, the variance of scores increased to 214.69. The measure of reliability paralleled the change in variance, with .7071 for the graduating students, .3016 for the beginning students, and .8860 for the combined groups. This last value of .8860 was expected to be more than the reliability for either group, as the combination of known competent students with known non-competent students forced an increase in the overall variability in scores.

Classical Analysis

In that the test items were given to a particular group of students who had been deemed by their respective schools to be eligible to write registration examinations in nursing, it was assumed that the majority of students would correctly answer items relevant to nursing knowledge. Given a 10 percent chance of error, this would leave the expected difficulty level for these items at 0.900. Point biserial correlations were not expected to be particularly revealing as the maximum obtainable value for an item with a 0.900 difficulty level would be 0.580 (Lord and Novick, 1968, p. 340). In fact, with little

variance among the examinees, there would also be a high probability of having negative values for the point biserial correlations. The distractor discriminating powers for the items would also show minimal difference if 90 percent of the examinees answered an item correctly, and, as with the point biserial correlations, there would be a tendency towards negative as well as positive results.

Given these expectations of how an item ought to perform, it was only possible to identify seemingly faultless items according to difficulty level. There were 16 items which had a difficulty level of between 0.900 and 1.000. All of these items were four distractor items. For these items, the point biserial correlations ranged from -0.015 to 0.639, with one of them being a negative value. The one point biserial correlation of 0.639 was greater than the expected maximum reported by Lord and Novick (1968). It was, however, consistent with the work of Adams (1960), who indicated the presence of inflated point biserial correlations in positively skewed distributions. The histogram of the graduate student results shows this type of distribution (Figure II). The corresponding distractor

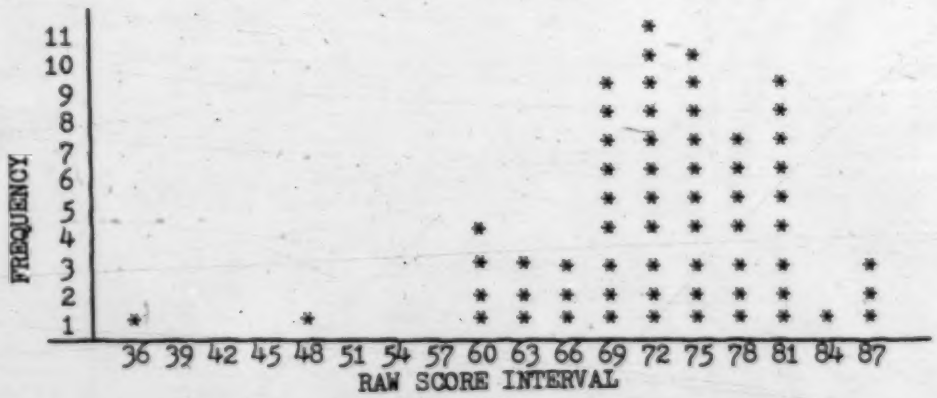

Figure II - Histogram of Group I Results

discriminating powers for the keyed answers to these items ranged from -0.01 to 0.22, with three of these being negative quantities.

By comparison to what was expected from this analysis, these 16 items might be termed the items which did not require revision. However, the decision was, in fact, based solely on the difficulty level, as the point biserial correlation and distractor discriminating powers did not aid in the decision making process. This analysis neither identified the reason for these items being easy (essential knowledge, poor construction, general non-nursing knowledge, etcetera); nor disclosed whether or not the item was easy for the group of non-competent examinees.

The item analysis data based on the performance of the graduating students is presented in Table II, with the results of the 16 criteria-meeting items indicated by asterisks.

Beginner - Graduate Difficulty Level Matrix

A crude sorting of item performance with the graduating, and then the beginning nursing students was carried out next. In order to develop the matrix, the difficulty levels of items were assessed to determine natural divisions into difficult, medium, and easy items. For graduating students, the difficult items had difficulty levels of 0.000 to 0.399, the medium items of 0.400 to 0.732, and the easy items of 0.733 to 1.000. For the beginning students, difficult items had difficulty levels ranging from 0.000 to 0.433, medium from 0.434 to 0.766, and easy items from 0.767 to 1.000. The matrix which was developed appears in Figure III.

Table II

Classical Item Analysis Data - Group I

ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.
001	.565	.301	.42	014	.710	.181	.27	026	.177	.110	.08
002	.581	-.048	.07	015	.452	-.010	.00	027	.677	.143	.02
003	.435	-.079	-.06	016	.403	.045	.17	028	.032	-.016	-.06
004*	.968	.265	.11	017	.435	.106	.10	029*	.968	.431	.11
005*	.952	.357	.11	018	.839	.437	.39	030	.177	.024	.02
006	.452	.178	.29	019*	.968	.338	.06	031	.581	.071	.12
007	.306	.039	-.03	020	.613	.021	.01	032	.645	.391	.48
008*	.935	.247	-.01	021	.629	.153	.19	033	.290	.049	.03
009*	.952	.229	-.01	022	.048	-.024	-.06	034	.645	.356	.43
010	.468	.254	.35	023	.532	-.143	-.17	035	.774	.069	.27
011	.677	.374	.43	024	.790	.147	.09	036	.645	.157	.19
012	.597	.280	.43	025	.645	.157	.19	037	.677	.288	.13
013	.597	.339	.55					038	.323	.296	.28

Note: * indicates item meeting criterion of difficulty level between 0.900 and 1.000

DIFF. = Difficulty Level

P.B.C. = Point Biserial Correlation

D.D.P. = Distractor Discriminating Power

Table II - continued

Classical Item Analysis Data - Group I

ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.
039*	1.000	.000	.00	054	.210	.119	.08	069	.484	.120	.10
040	.871	.231	.22	055	.581	.041	.01	070*	.903	-.015	-.01
041	.710	.105	.14	056	.177	.043	.08	071	.645	.081	.13
042*	.919	.326	.11	057	.500	-.108	-.17	072	.516	.294	.35
043	.484	.238	.35	058	.403	.187	.22	073	.323	.077	-.02
044	.726	.364	.37	059	.194	-.015	.02	074	.887	.092	.22
045	.161	.161	.13	060	.145	-.034	.08	075	.452	.042	.10
046	.323	.359	.45	061	.629	-.006	-.05	076	.645	.330	.38
047	.516	-.036	-.05	062	.855	-.065	-.01	077	.629	.248	.26
048	.710	.230	.21	063	.065	-.113	.01	078	.355	.103	.23
049	.855	.200	.10	064	.258	-.102	-.08	079	.484	.249	.41
050	.532	.261	.29	065	.645	.238	.43	080*	.903	.153	.04
051	.323	.088	.03	066	.113	.099	.13	081	.839	.083	.16
052	.000	.000	.00	067	.758	.005	.09	082	.532	.132	.13
053	.532	-.052	-.10	068	.097	.232	.19	083	.710	.137	.20

Table II - continued

Classical Item Analysis Data - Group I

ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.
084	.548	.146	.19	099	.435	.213	.17	114	.645	.115	.08
085	.565	.042	.07	100	.581	.316	.42	115	.532	.408	.48
086	.629	.301	.44	101	.677	.390	.36	116	.516	.331	.47
087	.452	.034	.05	102	.661	.309	.13	117*	.919	.393	.22
088	.806	.122	.21	103	.435	.128	.29	118	.532	.029	.13
089	.774	.582	.56	104	.726	.081	-.17	119	.855	.341	.10
090	.129	.065	.02	105	.194	-.159	-.10	120	.129	.054	.01
091	.887	.555	.28	106*	.919	.521	.22	121	.065	-.128	-.43
092	.581	.138	.01	107	.339	.253	.33	122	.371	.276	.34
093	.548	.183	.01	108	.726	.142	-.04	123*	.952	.639	.17
094	.855	.247	-.11	109	.194	-.006	.03	124	.484	.219	.12
095	.758	.090	.15	110	.726	.126	.08	125	.726	.426	.43
096	.145	.102	.08	111	.161	.226	.08	126	.419	.356	.40
097	.339	.237	.23	112	.274	-.011	-.20	127	.548	.153	.19
098*	.903	.252	.17	113*	.919	.527	.22	128*	.935	.366	.10

GROUP I - GRADUATING STUDENTS

GROUP II - BEGINNING STUDENTS	DIFFICULT .000 to .399	MEDIUM .400 to .732	EASY .733 to 1.000
DIFFICULT .000 to .433	A	B	C
MEDIUM .434 to .766	D	E	F
EASY .767 to 1.000	G	H	I

Figure III - Matrix of Difficulty Levels for Graduating and Beginning Students

Items which would not need revision because they distinguished between competent and non-competent students were those falling into cell C. These items were easy for the graduating students and difficult for the beginning students. Items became progressively less discriminating as they fell into cells B and F, then into cells A, E, and I. Items falling into A, E and I cells were not valuable in determining competency as they showed little difference between graduating and beginning students. Items which fell into cells D, H, and G increased in discriminating ability, but discriminated in a negative direction. That is, the beginning students found these items easier than the graduating students. Therefore, items which fell into these cells were suspect and in need of revision.

The matrix sorting was conducted solely on the basis of a comparison of the difficulty levels of items tested on Group I, the graduating students, and Group II, the beginning students. No

comparison of point biserial correlations or distractor discriminating power was done, as these would be subjective comparisons, since little variance was evident in the testing of either Group I or Group II.

In total, ten items fell into the criterion cell. The distribution of all items is depicted in Figure IV.

Comparison of matrix sorting to classical analysis. The matrix sorting of items by comparison of difficulty levels revealed only one item (#39) which also met the criterion of a difficulty level of between 0.900 and 1.000 when tested on Group I. The remaining items when tested on the graduating students had difficulty levels ranging from 0.750 to 0.890. Of the other 15 items which had difficulty levels of 0.900 to 1.000 on the graduating students, nine were found to have difficulties of 0.434 to 0.766, and six with difficulties of 0.767 to 1.000 when tested on the beginning students.

The matrix method was determined to be more accurate in discriminating between those items which differentiated between competent and non-competent students as it eliminated those items which were easily answered by beginning students as well as graduating students. However neither method was substantiated by information on point biserial correlations or distractor discriminating powers.

Comparison of Actual and Ideal Performance for an Item

The third method of analyzing an item was based on the same principle as the matrix sorting, that is, the comparison of the performance of an item when tested on competent and on non-competent students. However, as the results were grouped together, the test

GROUP I - GRADUATING STUDENTS

	DIFFICULT 0.000 to 0.399	MEDIUM 0.400 to 0.732	EASY 0.733 to 1.000
DIFFICULT 0.000 to 0.433	7, 22, 26, 28, 33, 38, 45, 46, 51, 52, 54, 56, 60, 63, 64, 66, 68, 73, 78, 90, 96, 97, 105, 107, 109, 111, 112, 120, 121, 122. Total = 30	1, 2, 3, 6, 11, 12, 13, 14, 15, 16, 17, 20, 21, 23, 27, 32, 34, 37, 41, 43, 44, 47, 48, 50, 55, 57, 61, 69, 71, 72, 75, 76, 79, 82, 84, 85, 86, 87, 93, 99, 100, 103, 110, 115, 116, 118, 124, 125, 126. Total = 49	18, 35, 39, 67, 74 88, 89, 91, 95, 119. Total = 10 CRITERION CELL
MEDIUM 0.434 to 0.766	30, 59. Total = 2	10, 25, 31, 36, 53, 58, 65, 77, 83, 92, 101, 102, 104, 108, 114, 127. Total = 16	4, 5, 19, 24, 40, 42, 49, 62, 70, 81, 94, 106, 117, 123, 128. Total = 15
EASY 0.767 to 1.000	Total = 0	Total = 0	8, 9, 29, 80, 98, 113. Total = 6

Figure IV - Matrix Sorting of Item Difficulty Levels

group of students became heterogeneous. This created an increase in the variance, resulting in the possibility of comparing point biserial correlations and distractor discriminating powers.

First, the actual difficulty levels of the items tested on the combined groups of students were compared using a significant difference of proportions test, to the ideal difficulty levels expected for the items. At a probability level of $p = .01$ there was no significant difference between the actual and ideal item analysis for 60 items. Of these 60 items, 13 showed no significant difference between actual and ideal point biserial correlations at a probability level of $p = .01$. The comparison of the actual and ideal distractor discriminating powers revealed only eight of the items which showed no significant difference at the probability level of $p = .01$.

Table III shows the significant difference values of difficulty levels, point biserial correlations, and distractor discriminating powers, with those values showing no significant difference at $p = .01$ indicated with an asterisk. As is shown in this table, eight items met all three criteria. These items are indicated with a double asterisk.

Comparison of Ideal - Actual analysis to Matrix Sorting and Classical analysis. By means of the comparison of ideal and actual analyses, all ten items identified as meeting criterion by the matrix sorting method were also identified. However, only seven of these items met the criterion of having no significant difference in point biserial correlations, and six with no significant difference in distractor discriminating powers. An additional two items not identified in the matrix met the criteria of difficulty levels, point biserial correlations and distractor discriminating powers; and an

Table III

Significant Difference Values (z)

Between Ideal and Actual Item Analysis Results

ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.
001	3.740	1.626*	3.844	017	3.361	4.705	9.712
002	1.971*	4.043	7.339	018**	0.921*	0.034*	0.191*
003	5.001	3.905	8.128	019	-4.297	2.323*	5.228
004	-4.163	2.555*	5.671	020	0.921*	4.198	8.909
005	-5.868	4.172	10.674	021	1.232*	3.690	6.742
006	3.740	4.103	9.715	022	6.662	8.422	20.261
007	3.271	6.477	18.450	023	0.644*	6.486	13.388
008	-7.029	4.981	13.570	024	-2.092*	3.768	8.665
009	-7.185	4.611	12.557	025	-0.331*	4.361	9.192
010	-1.612*	7.570	25.100	026	6.947	5.746	13.616
011	0.348*	3.157	6.578	027	0.122*	3.613	6.854
012	1.392*	3.666	6.119	028	8.005	8.077	20.178
013	2.993	2.099*	4.742	029	-8.116	4.146	13.582
014	0.348*	3.535	7.605	030	1.806*	9.566	33.141
015	2.394*	6.056	12.454	031	-0.852*	6.254	13.808
016	3.609	5.591	11.006	032	2.447*	1.789*	4.506

* indicates no significant difference at p = .01

** indicates item meeting criteria of difficulty level, point biserial correlation, and distractor discriminating power.

DIFF. = Difficulty Level

P.B.C. = Point Biserial Correlation

D.D.P. = Distractor Discriminating Power

Table III - continued

Significant Difference Values (z)

Between Ideal and Actual Item Analysis Results

ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.
033	4.620	6.004	14.533	057	3.490	4.559	8.155
034	0.644*	3.665	6.718	058	1.165*	6.323	18.055
035	-0.680*	3.432	7.141	059	2.036*	7.854	30.172
036	-0.802*	4.490	9.196	060	7.713	5.867	15.108
037	2.093*	1.652*	3.847	061	2.553*	2.624	4.267
038	4.867	4.903	11.530	062	-1.968*	2.994	4.743
039**	-1.860*	0.559*	0.000*	063	7.999	7.828	18.551
040	-4.945	4.422	10.369	064	6.234	5.505	11.832
041	-0.852*	4.361	7.693	065	-1.258*	5.127	13.141
042	-5.757	4.499	12.215	066	7.083	6.873	16.896
043	3.236	4.103	8.900	067	2.093*	1.385*	2.794
044	0.469*	2.931*	4.737	068	6.234	7.888	21.644
045	5.904	6.357	16.274	069	1.619*	6.194	12.854
046	6.044	3.561	10.616	070	-5.084	4.637	10.662
047	0.522*	6.194	13.390	071	1.165*	3.716	6.573
048	0.469*	3.157	7.605	072	3.036	3.518	7.334
049	-3.537	3.630	7.119	073	2.910	5.280	15.561
050	3.943	2.649	6.349	074**	0.348*	0.198*	0.383*
051	4.114	6.013	14.440	075	4.002	3.690	7.363
052	9.593	7.484	18.284	076	1.847*	2.030*	2.794
053	0.122*	6.649	17.668	077	-2.561*	6.400	18.439
054	6.167	5.944	16.684	078	6.167	3.191	8.604
055	0.921*	4.920	10.898	079	1.742*	4.637	10.530
056	5.006	7.854	22.181	080	-6.740	5.110	13.564

Table III - continued

Significant Difference Values (z)

Between Ideal and Actual Item Analysis Results

ITEM	DIFF.	P.B.C.	D.D.P.	ITEM	DIFF.	P.B.C.	D.D.P.
081	-3.077	4.447	8.440	105	6.309	6.151	16.122
082	1.452*	5.092	9.680	106**	-2.200*	1.178*	1.563*
083	-4.430	6.822	16.136	107	5.134	4.172	10.991
084	3.634	2.710	5.178	108	-2.092*	4.697	9.900
085	3.144	3.329	5.877	109	4.867	7.183	23.534
086	2.553*	2.503*	6.843	110	0.070*	3.665	6.982
087	1.574*	6.116	12.529	111	4.228	8.129	23.204
088**	1.514*	0.774*	0.771*	112	6.512	4.619	9.906
089**	1.847*	-0.610*	0.966*	113	-6.893	4.723	11.990
090	5.407	8.043	20.288	114	-0.974*	5.049	8.423
091	-0.924*	1.617*	4.926	115	2.785	2.658	6.593
092	-1.736*	6.417	19.057	116	3.121	3.312	7.915
093	0.921*	4.809	11.431	117	-4.047	2.658	5.205
094	-2.798	3.037	6.836	118	2.910	3.948	8.628
095	0.122*	2.787	5.192	119**	-0.226*	0.895*	0.771*
096	5.672	6.839	15.597	120	7.276	5.996	16.258
097	4.357	4.989	13.056	121	6.947	4.869	20.970
098	-5.757	4.284	11.273	122	4.244	4.250	9.763
099	2.093*	5.308	11.483	123	-3.668	1.901*	4.065
100	1.574*	3.716	6.845	124	3.144	3.974	10.231
101	-1.673*	4.869	10.236	125**	0.871*	1.566*	2.484*
102	0.122*	3.768	7.351	126	2.394*	5.583	13.055
103	4.734	3.277	8.132	127	-2.326*	7.338	23.876
104	-3.782	5.858	13.621	128	-3.668	2.443*	6.115

additional five met the difficulty level and point biserial correlation criteria.

This method of testing for significant difference gave more explicit information as to needed improvements than either the classical analysis or matrix sorting. It eliminated those easy items identified by the classical analysis, which did not discriminate between beginning and graduating students. It also used more precise criteria than the matrix sorting to identify items which discriminated between competent and non-competent students.

The comparison of the ideal and actual difficulty levels alone did not eliminate those items which the beginning students found easier than the graduating students. Howver, these items did not meet the point biserial correlation criterion.

The comparison of both point biserial correlations and distractor discriminating powers increased the usefulness of the analysis method, as these computations identified areas requiring revision. The other two analysis methods did not identify specific problem areas, and therefore were not as useful for item revision. A comparison of the items meeting the criteria of the three analysis techniques is presented in Table IV.

Table IV

Item by Item Comparison of Three Analysis Methods

ITEM	CLASSICAL ANALYSIS	MATRIX SORTING	ACTUAL - IDEAL COMPARISON		
			DIFF.	P.B.C.	D.D.P.
001				*	
002			*		
003					
004	*			*	
005	*				
006					
007					
008	*				
009	*				
010			*		
011			*		
012			*		
013				*	
014			*		
015			*		
016					
017					
018		*	*	*	*
019	*			*	
020			*		

Note:

* indicates item meeting criterion

DIFF. = Difficulty Level

P.B.C. = Point Biserial Correlation

D.D.P. = Distractor Discriminating Power

Table IV - continued

Item by Item Comparison of Three Analysis Methods

ITEM	CLASSICAL ANALYSIS	MATRIX SORTING	ACTUAL - IDEAL COMPARISON		
			DIFF.	P.B.C.	D.D.P.
021			*		
022					
023			*		
024			*		
025			*		
026					
027			*		
028					
029	*				
030			*		
031			*		
032			*	*	
033					
034			*		
035		*	*		
036			*		
037			*	*	
038					
039	*	*	*	*	*
040					
041			*		
042	*				
043					
044			*	*	
045					

Table IV - continued

Item by Item Comparison of Three Analysis Methods

ITEM	CLASSICAL ANALYSIS	MATRIX SORTING	ACTUAL - IDEAL COMPARISON		
			DIFF.	P.B.C.	D.D.P.
046					
047			*		
048			*		
049					
050					
051					
052					
053			*		
054					
055			*		
056					
057					
058			*		
059			*		
060					
061			*		
062			*		
063					
064					
065			*		
066					
067		*	*	*	
068					
069			*		
070	*				

Table IV - continued

Item by Item Comparison of Three Analysis Methods

ITEM	CLASSICAL ANALYSIS	MATRIX SORTING	ACTUAL - IDEAL COMPARISON		
			DIFF.	P.B.C.	D.D.P.
071			*		
072					
073					
074		*	*	*	*
075					
076			*	*	
077			*		
078					
079			*		
080	*				
081					
082			*		
083					
084					
085					
086			*	*	
087			*		
088		*	*	*	*
089		*	*	*	*
090					
091		*	*	*	
092					
093			*		
094					
095		*	*		

Table IV - continued

Item by Item Comparison of Three Analysis Methods

ITEM	CLASSICAL ANALYSIS	MATRIX SORTING	ACTUAL - IDEAL COMPARISON		
			DIFF.	P.B.C.	D.D.P.
096					
097					
098	*				
099			*		
100			*		
101			*		
102			*		
103					
104					
105					
106	*		*	*	*
107					
108			*		
109					
110			*		
111					
112					
113	*				
114			*		
115					
116					
117	*				
118					
119		*	*	*	*
120					

Table IV - continued

Item by Item Comparison of Three Analysis Methods

ITEM	CLASSICAL ANALYSIS	MATRIX SORTING	ACTUAL - IDEAL COMPARISON		
			DIFF.	P.B.C.	D.D.P.
121					
122					
123	*			*	
124					
125			*	*	*
126			*		
127			*		
128	*			*	

Summary

Although the calculation of ideal analysis values and computation of significant differences between actual and ideal difficulty levels, point biserial correlations, and distractor discriminating powers required additional computational time, the method gave the most precise data and clearest guidelines for the revision of items.

The classical analysis did little to identify items which would discriminate between competent and non-competent students. It served only to identify the items with which the competent students are familiar. The matrix sorting used subjective cut-off points and gave none of the additional information provided by point biserial correlations or distractor discriminating powers as to where changes in the item were required in order to improve it.

The following chapter contains a discussion of the usefulness of the information provided by the adapted item analysis methodology, problems inherent in the administration of the item analysis, problems in the analysis itself, and suggestions for further study related to this classical analysis adaptation.

V. Conclusions and Recommendations

An adaptation of the classical item analysis was used to increase the utility of information about test items which were being used to evaluate nursing students. Following a technique suggested by Haladyna (1974), Kosecoff and Fink (1976), and Popham and Husek (1969); a forced variance was introduced by combining beginning and graduating students in the examinee group. With the characteristics of the group known, an expected item analysis result for the difficulty level, point biserial correlation, and distractor discriminating power was calculated. These ideals were compared to the actual item analysis values for each item, with the result that information was available as to how well the items were able to be used to differentiate between groups of competent and non-competent nursing students. Additional information which could be used in the revision of items, specifically in terms of the difficulty levels, point biserial correlations, and distractor discrminating powers was also provided through the adapted methodology.

Examples of the use of the information provided by this item analysis methodology are provided in this chapter. The problems inherent in the administration of the methodology are examined and additionally, suggestions for future research are given.

Analysis of Specific Items

The analysis of specific items is presented by item groups; those which met the criteria on all three domains of difficulty level, point biserial correlation, and distractor discriminating power; those

which met the criteria of difficulty level and point biserial correlation; those which met the criterion of difficulty level alone; those which met the criterion of point biserial correlation; and those which met none of the three criteria.

Items meeting criteria for difficulty level, point biserial correlation, and distractor discriminating power. Eight items met the criteria in all three categories. That is, for these items, there was no significant difference between the actual item analysis results, and the ideal results expected for the difficulty level, point biserial correlation, and distractor discriminating power. These items, numbers 18, 39, 74, 88, 89, 106, 119 and 125 are acceptable items for use in differentiating between competent and non-competent nursing students. These items should be identified as such in the item bank so that they can be readily retrieved for examinations used to identify competency in nursing.

Items meeting criteria for difficulty level and point biserial correlation. Seven items, numbers 32, 37, 44, 67, 76, 86 and 91 met the criteria for both difficulty level and point biserial correlation, but failed to meet the criteria for distractor discriminating power. Although these items discriminated between competent and non-competent students, the keyed distractor did not perform as well as it should have.

Each of the above items was a four distractor item. Therefore the ideal distractor discriminating power for the keyed answer would have been 0.65. The actual distractor discriminating powers ranged from 0.33 to 0.51. Each of the distractors should be reviewed to determine why non-competent students were choosing the keyed distractor, or why competent students were choosing the alternatives. Based on this

review, the item distractors should be revised, and the item retested.

Items meeting criterion for difficulty level. Forty-five items met the criterion for difficulty level, but failed to meet either the point biserial correlation or the distractor discriminating power criteria. These items, if reviewed solely on the basis of difficulty level might have been accepted as items which discriminated between competent and non-competent students. However, in each case, either fewer competent students chose the correct distractor than would have in the ideal case, or more non-competent students chose the correct distractor than would have been expected. These items, numbers 2, 10, 11, 12, 14, 15, 20, 21, 23, 24, 25, 27, 30, 31, 34, 35, 36, 41, 47, 48, 53, 55, 58, 59, 61, 62, 65, 69, 71, 77, 79, 82, 87, 92, 93, 95, 99, 100, 101, 102, 108, 110, 114, 126 and 127 should be reviewed as to their content and structure to determine possible faults. They should be reviewed to determine why alternatives were attractive to the competent students, and why the keyed distractor was attractive to the non-competent students. Revisions should be made accordingly and the items retested.

Items meeting criterion for point biserial correlation. Six items, numbers 1, 4, 13, 19, 123 and 128 met only the criterion for point biserial correlation, but failed to meet the criteria for difficulty level and distractor discriminating power. These six items were able to discriminate between competent and non-competent students. However, they were either too easy or too hard for both groups of students. In reviewing these items, the key and other distractors should be studied closely, to ensure that the item was correctly keyed, and that

no other distractor is a possible key. Following necessary revisions, the items should be retested.

Items meeting no criteria. The largest group of items, 61 in number met none of the criteria. Initial revisions to these items could be based on the difficulty levels of the items. These fall into two categories; those items which proved to be too easy, and those which were too difficult for the students in either group.

The 16 items, numbers 5, 8, 9, 29, 40, 42, 49, 70, 80, 81, 83, 94, 98, 104, 113 and 117, with difficulty levels above .650 proved to be easy for both competent and non-competent students. These items should be reviewed for possible clues in the content or structure of the item. An item within this group may well test knowledge which is essential to nursing, but which is also known to a variety of other non-nursing persons. Items of this type should be identified as such by nursing experts, and should not be used to discriminate between those who are competent and those who are not competent in nursing.

The remaining 45 items proved difficult for both competent and non-competent students. These items should be reviewed by nursing experts to establish whether or not the content is essential to nursing. The structure of the item should also be reviewed to determine whether or not any of the alternative distractors was misleading. The keyed distractors should also be checked for accuracy. As with the other items, the revised items should be retested and reanalyzed.

Appendix II contains the item analysis according to the above five categories.

Conclusions

Although there are difficulties in obtaining subject groups for the administration of the test examination, and in the computation of ideal results, the item analysis information provided by the adapted methodology is more readily usable than that provided by a classical item analysis procedure. The information provided by the classical technique provided some false positive results as it did not take the non-competent group of students into consideration. In such cases, the item appeared on the basis of difficulty level to meet the set criterion, but in fact was easy for the non-competent students as well.

Information provided as to point biserial correlation and distractor discriminating powers was useful as a guide to item revisions. By giving the information in terms of difference from ideal values, items requiring revisions were accurately pinpointed, decreasing the subjectivity of revisions based on the classical analysis results.

More accurate information as to the discriminating ability of the item was provided by the matrix sorting of the analysis results for the competent and for the non-competent students than was provided by the classical analysis method. However, the cutting points for the matrix were subjectively assigned, and information as to point biserial correlations and distractor discriminating powers for the combined groups was not available to aid in revisions. Therefore, the adapted methodology proved more useful than the matrix sorting.

Discussion and Implications

Retesting of items by this item analysis procedure would have

inherent in it some of the difficulties which arose in the original procedure. The original groups of students used for the study were from only two schools of nursing in Alberta, and thus were from a very limited subsection of the population. However, if numbers of participating schools were increased, difficulties would arise in administrative costs due to the necessity of travel. Timetabling of non-essential examinations would also present some difficulties, and would require the full cooperation of the schools involved.

As the examination is non-essential in nature; that is, it does not contribute to the evaluation of the student, motivation of students is difficult to achieve. This may have contributed to the relatively low marks in the competent group. Adding a few test items to a school-set examination rather than having a complete examination of test items might increase the motivation of the students in answering the test items. Retesting of the eight items which met all three criteria of difficulty level, point biserial correlation, and distractor discriminating power within the context of another group of items would increase the reliability of the results attained in the present study.

Difficulties in analysis procedure. The adapted analysis procedure, because it includes more calculations than the classical technique, takes longer to provide information for item revisions. Some of the additional computations as to ideal values for difficulty levels, point biserial correlations, and distractor discriminating powers can be worked out in advance of the item analysis procedure.

The required revisions to the TEST04 program can be done from a

histogram of the combined group results and do not require the entire TEST04 analysis procedure. The classical analysis could then be dropped in favour of the adapted analysis procedure.

Computations comparing the actual and ideal analysis values could be speeded up by developing a computer program which could work from the adapted TEST04 program, with input of the expected ideal values. Streamlining of these procedures for the use of the computer would ease the analysis of the test items, and increase the utility of the adapted procedure. As it stands, the adapted methodology may appear too cumbersome to be used by those people who are charged with the task of revising test items.

Suggestions for Further Study

As the study used only a small proportion of the available items in the University of Alberta Hospital School of Nursing item bank, it would be useful to repeat the study using other groups of items from the bank. Additionally, items used in this study should be revised and retested to determine whether or not the information provided by the adapted methodology assisted in the revision procedure. The eight criterion items should be retested within a different context of examination items in order to see if their classification as criterion-meeting items is reliable.

As this study used known groups of competent and non-competent students to analyze the test items, different groups of students could be used, as long as they met the specifications of competence and non-competence in the field of study being tested. By using

different groups of students within the school of nursing, test items could be identified for use at various levels within the school in order to assess competence at a specific level. Also, by using various levels of students within one school, it would be easier to add test items to established examinations. This would also make possible the comparison of various levels of competence rather than the dichotomy of competence and non-competence.

Items might also be arranged by subject type in order to decrease difficulties in the many transitions from subject to subject in the original study. This adaptation would increase the knowledge as to which subject areas needed the most revision in items. Information thus provided would assist in both the revision of, and the creation of new items, as items which met criterion could be used as guides for new items within the context area.

Information from the items used in this study could be reviewed in terms of the item type, that is situational or non-situational; single answer or multiple component; and four or five distractor. Information as to which types best discriminated between competent and non-competent students could be used as a guide for item writers. Possible sources of errors in the other items could also be listed and used to assist item writers to avoid these faults.

Development of a computer program which could compute ideal values, and significant differences between ideal and actual analysis results would increase the usefulness of the adapted methodology. A program of this type would alleviate those revising items of lengthy calculations which in themselves might deter instructors from using the adapted

methodology.

The actual, rather than the surmised usefulness of the adapted methodology could be studied by involving instructors in the revision of items, using the information provided by this technique.

Summary

An adaptation to the classical item analysis procedure was devised so that the information provided indicated more precisely which items were able to discriminate between competent and non-competent students. The technique involved a comparison of the known groups of competent and non-competent students in answering test items. The resultant item analysis results were compared to ideal analysis results which had been previously computed. Information thus provided was used to indicate the type of revisions which were needed in order to improve the items.

Although the method used presented some difficulties in enlisting the aid of volunteer students to write the examination, the procedure proved worthwhile. Suggestions were made as to variations in the examinee group, and as to the administration of the items as part of an evaluative examination in order to decrease administrative difficulties.

The adapted methodology appears to be a viable alternative to the classical analysis technique and should increase the efficiency of item revisions as well as accurately identify those items which do in fact differentiate between competent and non-competent students.

References

Adams, J. F. The effect of non-normally distributed criterion scores on item analysis techniques. Educational and Psychological Measurement, 1960, 20(2), 317-320.

Alker, H. A., Carlson, J. A., & Hermann, M. G. Multiple-choice questions and student characteristics. Journal of Educational Psychology, 1969, 60(3), 231-243.

American Psychological Association Committee on Test Standards. Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 1954, 51(2), 12-37.

Baker, F. B. An intersection of test score interpretation and item analysis. Journal of Educational Measurement, 1963, 1(1), 23-28.

Bloom, B. S. Recent developments in mastery learning. Educational Psychologist, 1973, 10(2), 53-57.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14(3), 277-289.

Cronbach, L. J. Test validation. In Thorndike, R. L. Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Diamond, J. J., & Evans, W. J. An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement. 1972, 9(2), 145-150.

Dressel, P. L. The role of external testing programs in education. Educational Record, 1964, 45(2), 161-166.

Dzubian, C. D., & Vickery, K. V. Criterion-referenced measurement: some recent developments. Educational Leadership, 1973, 30(5), 483-486.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs: Prentice-Hall, 1965.

Ebel, R. L. Criterion referenced measurements: limitations. School Review, 1971, 79(2), 282-288.

Esler, W. K., & Dzubian, C. D. Criterion referenced test: some advantages and disadvantages for science instruction. Science Education, 1974, 58(2), 171-174.

Feldt, L. S. The use of extreme groups to test for the presence of a relationship. Psychometrika, 1961, 26(3), 307-316.

Ferguson, G. A. Statistical analysis in psychology and education (4th ed.). Toronto: McGraw-Hill, 1976.

Glaser, R. Instructional technology and the measurement of learning outcomes: some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R. A criterion-referenced test. In Popham, W. J. (ed.). Criterion referenced measurement, an introduction. Englewood Cliffs: Educational Technology Publications, 1971.

Gulliksen, H. Theory of mental tests. New York: John Wiley & Sons, 1950.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11(2), 93-99.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10(3), 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, 1978, 48(1), 1-47.

Henrysson, S. Gathering, analyzing and using data on test items. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council of Education, 1971.

Hively, W. (ed.). Domain-referenced testing. Englewood Cliffs: Educational Technology Publications, 1974.

Hofmann, R. J. The concept of efficiency in item analysis. Educational and Psychological Measurement, 1975, 35, 621-640.

Hopkins, K. D. Response styles, chance and other irrelevant sources of variation on test performance. In Bracht, G., Hopkins, K., & Stanley, J. (eds.). Perspectives in educational and psychological measurement. Englewood Cliffs: Prentice-Hall, 1972.

Hubbard, J. P., & Clemans, W. V. Multiple choice examinations in medicine. A guide for examiner and examinee. Philadelphia: Lea and Febiger, 1961.

Hunt, D., & Randhawa, B. S. Perspectives and prospects of criterion referenced testing. Saskatchewan Journal of Educational Research and Development, 1976, 7(1), 3-14.

Jones, L. V. The nature of measurement. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Kelley, T. L. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 1939, 30, 17-24.

Klein, S. P. Evaluating tests in terms of the information they provide. Evaluation Comment, 1970, 2(2), 1-6.

Kosecoff, J., & Fink, A. The appropriateness of criterion-referenced tests for evaluation studies. Washington, D. C.: National Institute of Education, 1976. (ERIC Document Reproduction Service No. ED 135 841).

Krathwohl, D. R., & Payne, D. A. Defining and assessing educational objectives. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Lindquist, E. F. (ed.). Educational measurement. Washington, D. C.: American Council on Education, 1951.

Lippey, G. (ed.). Computer-assisted test construction. Englewood Cliffs: Educational Technology Publications, 1974.

Livingston, S. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9(1), 13-25.

Loevinger, J. Objective tests as instruments of psychological theory. Psychological Reports, 1957, 3, 1-18.

Lord, F. M. Problems in mental test theory arising from errors of measurement. In Mehrens, W. A., & Ebel, R. L. (eds.). Principles of educational and psychological measurement. Chicago: Rand McNally and Company, 1967.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. New York: Addison-Wesley, 1968.

Manning, W. H. The functions and uses of educational measurement. In Proceedings of the 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1970.

Millman, J. Criterion-referenced measurement. In Popham, W. J. (ed.). Evaluation in education: current practices. Berkely: McCutchin Publishers, 1974.

Oosterhof, A. C. Similarity of various item discrimination indices. Journal of Educational Measurement, 1976, 13(2), 145-149.

Popham, W. J. An approaching peril: cloud-referenced tests. Phi Delta Kappan, 1974, 55, 614-615.

Popham, W. J. Educational evaluation. Englewood Cliffs: Prentice-Hall, 1975.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6(1), 1-9.

Sax, G. Principles of educational measurement and evaluation. Belmont, California: Wadsworth Publishing Co., 1974.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45(1), 127-147.

Stanley, J. C. Reliability. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Strang, H. R. The effects of technical and unfamiliar options on guessing on multiple-choice test items. Journal of Educational Measurement, 1977, 14(3), 253-259.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15(2), 111-115.

Thorndike, R. L. Reliability. In Proceedings of the 1963 invitational conference on testing problems. Princeton: Educational Testing Service, 1964.

Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Tyler, R. A generalized technique for constructing achievement tests. Educational Research Bulletin, 1931, 10(8), 199-208.

Vaughn, K. W. Planning the objective test. In Lindquist, E. F. (ed.). Educational measurement. Washington, D. C.: American Council on Education, 1951.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems. Princeton: Educational Testing Service, 1968.

BIBLIOGRAPHY

Adams, J. F. The effect of non-normally distributed criterion scores on item analysis technique. Educational and Psychological Measurement, 1960, 20(2), 317-320.

Adkins, D. C. Measurement in relation to educational process. Educational and Psychological Measurement, 1958, 18(2), 221-240.

Alker, H. A., Carlson, J. A., & Hermann, M. G. Multiple-choice questions and student characteristics. Journal of Educational Psychology, 1969, 60(3), 231-243.

American Psychological Association Committee on Test Standards. Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 1954, 51(2), 12-37.

Anastasi, A. (ed.). Invitational conference on testing problems. Testing problems in perspective; 25th anniversary volume of topical readings from the Invitational Conference on Testing Problems. Washington, D. C.: American Council on Education, 1966.

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Astin, A. W. Criterion-centered research. Educational and Psychological Measurement, 1964, 24(4), 807-822.

Baker, F. B. An intersection of test score interpretation and item analysis. Journal of Educational Measurement, 1963, 1(1), 23-28.

Baker, F. B. Advances in item analysis. Review of Educational Research, 1977, 47(1), 151-178.

Block, J. H. Criterion-referenced measurements: potential. School Review, 1971, 79(2), 289-298.

Bloom, B. S. Recent developments in mastery learning. Educational Psychologist, 1973, 10(2), 53-57.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw Hill Book Co., 1971.

Bock, R. D., & Wood, R. Test theory. Annual Review of Psychology, 1971, 22, 193-224.

Bracht, G., Hopkins, K., & Stanley, J. (eds.). Perspectives in educational and psychological measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14(3), 277-289.

Bridgman, C. S. The relation of the upper - lower item discrimination index, D, to the bivariate normal correlation coefficient. Educational and Psychological Measurement, 1964, 24(1), 85-90.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. Educational Horizons, 1970, 48, 71-80.

Coffman, W. E. Concepts of achievement and proficiency. In Proceedings of the 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1970.

Cox, R. C. Evaluative aspects of criterion-referenced measures. In Popham, W. J. (ed.). Criterion referenced measurement, an introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper Publishing Co., 1970.

Cronbach, L. J. Validation of educational measures. In Proceedings of the 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1970.

Cronbach, L. J. Test validation. In Thorndike, R. L. Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Denney, C. There is more to a test pool than data collection. Educational Technology, 1973, 13, 19-20.

Diamond, J. J., & Evans, W. J. An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 1972, 9(2), 145-150.

Dressel, P. L. The role of external testing programs in education. Educational Record, 1964, 45(2), 161-166.

Dzubian, C. D., & Vickery, K. V. Criterion-referenced measurement; some recent developments. Educational Leadership, 1973, 30(5), 483-486.

Ebel, R. L. Must all tests be valid? American Psychologist, 1961, 16(10), 640-647.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N. J.: Prentice-Hall, 1965.

Ebel, R. L. The relation of item discrimination to test reliability. Journal of Educational Measurement, 1967, 4(3), 125-128.

Ebel, R. L. Criterion-referenced measurements: limitations. School Review, 1971, 79(2), 282-288.

Engelhart, M. D. A comparison of several item discrimination indices. Journal of Educational Measurement, 1965, 2(1), 69-76.

Esler, W. K., & Dzubian, C. D. Criterion referenced test: some advantages and disadvantages for science instruction. Science Education, 1974, 58(2), 171-174.

Feldt, L. S. The use of extreme groups to test for the presence of a relationship. Psychometrika, 1961, 26(3), 307-316.

Ferguson, G. A. Statistical analysis in psychology and education (4th ed.). Toronto: McGraw Hill Book Co., 1976.

Findley, W. G. A rationale for evaluation of item discrimination statistics. Educational and Psychological Measurement, 1956, 16(2), 175-180.

Flanagan, J. C. Units, scores and norms. In Lindquist, E. T. (ed.). Educational measurement. Washington, D. C.: American Council on Education, 1951.

Glaser, R. Instructional technology and measurement of learning outcomes: some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R. A criterion-referenced test. In Popham, W. J. (ed.). Criterion referenced measurement. An introduction. Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

Glaser, R., & Klaus, D. J. Proficiency measurement: assessing human performance. In Gagne, R. (ed.). Psychological principles in system development. New York: Holt, Rinehart & Winston, 1962.

Gronlund, N. E. Measurement and evaluation in teaching (3rd ed.). New York: MacMillan, 1976.

Gulliksen, H. Theory of mental tests. New York: John Wiley & Sons, 1950.

Guttman, L. Integration of test design and analysis. In Proceedings of the 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1970.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11(2), 93-99.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44(4), 371-400.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14(2), 75-96.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10(3), 159-170.

Hambleton, R. K., Swaminathan, H., & Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In de Gruijter, D. N. M., & van der Kamp, L. J. T. (eds.). Advances in psychological and educational measurement. New York: Wiley & Sons, 1976.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, 1978, 48(1), 1-47.

Henrysson, S. Gathering, analyzing and using data on test items. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Hieronymous, A. Today's testing: what do we know how to do? In Proceedings of the 1971 invitational conference on testing problems, Princeton: Educational Testing Service, 1972.

Hively, W. (ed.). Domain-referenced testing. Englewood Cliffs, N. J.: Educational Technology Publications, 1974.

Hofmann, R. J. The concept of efficiency in item analysis. Educational and Psychological Measurement, 1975, 35, 621-640.

Hopkins, K. D. Response styles, chance and other irrelevant sources of variation in test performance. In Bracht, G., Hopkins, K., & Stanley, J. (eds.). Perspectives in educational and psychological measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Hubbard, J. P. Measuring medical education. The tests and test procedures of the National Board of Medical Examiners. Philadelphia: Lea & Febiger, 1971.

Hubbard, J. P., & Clemans, W. V. Multiple choice examinations in medicine. A guide for examiner and examinee. Philadelphia: Lea & Febiger, 1961.

Hughes, H. H., & Trimble, W. E. The use of complex alternatives in multiple choice items. Educational and Psychological Measurement, 1965, 25, 117-126.

Hunt, D., & Randhawa, B. S. Perspectives and prospects of criterion referenced testing. Saskatchewan Journal of Educational Research and Development, 1976, 7(1), 3-14.

Jones, L. V. The nature of measurement. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Keats, J. A. Test theory. Annual Review of Psychology, 1967, 18, 217-238.

Kelley, T. L. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 1939, 30, 17-24.

Klein, S. P. Evaluating tests in terms of the information they provide. Evaluation Comment, 1970, 2(2), 1-6.

Kosecoff, J., & Fink, A. The appropriateness of criterion-referenced tests for evaluation studies. Washington, D. C.: National Institute of Education, 1976. (ERIC Document Reproduction Service No. ED 135 841).

Krathwohl, D. R., & Payne, D. A. Defining and assessing educational objectives. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lange, A., Lehmann, I. J., & Mehrens, W. A. Using item analysis to improve tests. Journal of Educational Measurement, 1967, 4, 65-68.

Levine, H.G., McGuire, C.H., & Nattress, L. W. Jr. The validity of multiple choice achievement tests as measures of competence in medicine. American Educational Research Journal, 1970, 7(1), 69-82.

Lindeman, R. H. Educational measurement. Glenview, Illinois: Scott, Foresman & Co., 1967.

Linden, J. D., & Linden, K. W. Tests on trial: guidance monograph series III: testing. Boston: Houghton Mifflin Co., 1968.

Lindquist, E. F. (ed.). Educational measurement. Washington, D. C.: American Council on Education, 1951.

Lippey, G. Computer assisted test construction. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Livingston, S. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9(1), 13-25.

Loevinger, J. Objective tests as instruments of psychological theory. Psychological Reports, 1957, 3, 1-18.

Lord, F. M. The relation of the reliability of multiple choice tests to the distribution of item difficulties. Psychometrika, 1952, 17, 181-194.

Lord, F. M. The effect of random guessing on test validity. Educational and Psychological Measurement, 1964, 24, 745-747.

Lord, F. M. Problems in mental test theory arising from errors of measurement. In Mehrens, W. A., & Ebel, R. L. (eds.). Principles of educational and psychological measurement. Chicago: Rand McNally & Co., 1967.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. New York: Addison-Wesley, 1968.

Madaus, G. F. The predictive validity of the NLN pre-nursing and guidance examination for different criteria of success in a three-year diploma program. Educational and Psychological Measurement, 1966, 26, 431-437.

Mager, R. F. Measuring instructional intent: or got a match? Belmont: California: Fearon Publishers, 1973.

Manning, W. H. The function and uses of educational measurement. In Proceedings of the 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1970.

McClelland, D. C. Testing for competence rather than for "intelligence". American Psychologist, 1973, 28(1), 1-14.

McGuire, C. H. An evaluation model for professional education: medical education. In Proceedings of the 1967 invitational conference on testing problems. Princeton: Educational Testing Service, 1968.

McKenna, B., Taylor, E., Darehshori, C., Engel, B., & Quinto, F. Testing. Today's Education, 1977, 66(2), 34-55.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: views regarding mastery and standard setting. Review of Educational Research, 1976, 46(1), 133-158.

Millman, J. Criterion-referenced measurement. In Popham, W. J. (ed.). Evaluation in education: current practices. Berkely: McCutchin Publishing Co., 1974.

Myers, C. T. The relationship between item difficulty and test validity and reliability. Educational and Psychological Measurement, 1962, 22, 565-571.

National Association of Secondary School Principals. Competency tests and graduation requirements. Reston, Virginia: National Association of Secondary School Principals, 1976.

Noll, V. H. Testing under fire. In Proceedings of the 1964 invitational conference on testing problems. Princeton: Educational Testing Service, 1965.

O'Connor, E. F. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42, 73-97.

Oosterhof, A. C. Similarity of various item discrimination indices. Journal of Educational Measurement, 1976, 13(2), 145-149.

Popham, W. J. An approaching peril: cloud-referenced tests. Phi Delta Kappan, 1974, 55, 614-615.

Popham, W. J. Educational evaluation. Englewood Cliffs, N. J.: Prentice-Hall, 1965.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6(1), 1-9.

Pyrczak, F. Validity of discrimination index as a measure of item quality. Journal of Educational Measurement, 1973, 10(3), 227-231.

Quirk, T. J. Some measurement issues in competency-based teacher education. Phi Delta Kappan, 1974, 55(5), 316-319.

Richardson, M. W., & Stalnaker, J. M. Comments on achievement examinations. Journal of Educational Research, 1935, 28, 425-432.

Sax, G. Principles of educational measurement and evaluation. Belmont, California: Wadsworth Publishing Co., 1974.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45(1), 127-147.

Stanley, J. C. Reliability. In Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Strang, H. R. The effects of technical and unfamiliar options on guessing on multiple-choice test items. Journal of Educational Measurement, 1977, 14(3), 253-259.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15(2), 111-115.

Thorndike, R. L. Reliability. In Proceedings of the 1963 invitational conference on testing problems. Princeton: Educational Testing Service, 1964.

Thorndike, R. L. (ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Tyler, R. A generalized technique for constructing achievement tests. Educational Research Bulletin, 1931, 10, 199-208.

Tyler, R. W., & Wolf, R. M. (eds.). Crucial issues in testing. Berkely, California: McCutchan, 1974.

Vaughn, K.W. Planning the objective test. In Lindquist, E. F. (ed.). Educational measurement. Washington, D. C.: American Council on Education, 1951.

Wahlstrom, M. W. A factor analytic item selection procedure. Unpublished Master's thesis, University of Alberta, 1967.

Woodson, M. I. C. E. The issues of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11(1), 63-64.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems. Princeton: Educational Testing Service, 1968.

APPENDICES

APPENDIX I

TYPES OF EXAMINATION ITEMS

A. NON-SITUATIONAL SINGLE ANSWER:

All except which of the following statements are related to infantile colic?

a. It is more common in first born children.

b. It occurs most frequently between 5 p.m. and 12 a.m.

c. The parents of the infant are usually young and tense.

d. It is due to a congenital stricture not needing surgical correction.

B. NON-SITUATIONAL MULTIPLE COMPONENT:

Parents of a male child with genito-urinary problems need to understand that

1. Elective surgery should be performed before school age.
2. Rarely are the kidneys damaged by these problems.
3. All essential surgery should be performed before puberty.
4. Infections infrequently accompany these problems.

Select answer:

a. 1 and 2 are correct.

b. 1 and 3 are correct.

c. 1 and 4 are correct.

d. all of the above are correct.

SITUATION:

THE FOLLOWING STEM PERTAINS TO QUESTIONS C. AND D.

Mr. J. was admitted to hospital with a diagnosis of peripheral vascular disease. He exhibits signs and symptoms of dry gangrene in his right leg. He is scheduled for a mid-thigh amputation.

C. SITUATIONAL MULTIPLE COMPONENT:

In the preop preparation of Mr. J., the nurse should:

1. Make the patient aware of some of the difficulties he may encounter when he tries to use an artificial limb.
2. Explain that the patient should lie on his back.
3. Explain to Mr. J. that he may experience phantom limb pain.
4. Explain to Mr. J. that he will be expected to turn frequently.

Select answer:

a. 1, 2, and 3 are correct.

b. 1 and 3 are correct.

c. 2 and 4 are correct.

d. 4 is correct.

e. 1, 2, 3, and 4 are correct.

D. SITUATIONAL SINGLE ANSWER:

Mr. J. will probably be fitted for a prosthesis:

a. when the stump heals.

b. after he has mastered crutch walking.

c. when weight bearing is permitted.

d. when the pain has disappeared.

APPENDIX II

ADAPTED ITEM ANALYSIS RESULTS

NOTE: Items are grouped under the following categories:

1. Items meeting criteria in difficulty level, point biserial correlation, and distractor discriminating power.
2. Items meeting criteria of difficulty level and point biserial correlation.
3. Items meeting criterion of difficulty level.
4. Items meeting criterion of point biserial correlation.
5. Items meeting no criteria.
    a. difficulty level above .650.
    b. difficulty level below .360.

Difficulty levels (D. L.), point biserial correlations (P. B. C.) and distractor discriminating powers (D. D. P.) are reported for each item.

1. Items meeting criteria of difficulty level, point biserial correlation, and distractor discriminating power.

No revisions necessary.

ITEM	D.L.	P.B.C.	D.D.P.
018	.464	.638	.64
039	.623	.600	.65
074	.497	.627	.63
088	.430	.584	.61
089	.411	.680	.60
106	.642	.552	.59
119	.530	.574	.61
125	.437	.580	.57

2. Items meeting criteria of difficulty level and point biserial correlation.

Revisions to distractors needed to improve discriminating ability.

ITEM	D.L.	P.B.C.	D.D.P.
032	.377	.501	.43
037	.397	.513	.46
044	.490	.446	.42
067	.397	.535	.51
076	.411	.479	.51
086	.371	.436	.33
091	.570	.515	.44

3. Items meeting criterion of difficulty level.

Revisions needed in content and/or structure to increase appeal of keyed distractor to competent students, and decrease appeal of keyed distractor to non-competent students.

ITEM	D.L.	P.B.C.	D.D.P.	ITEM	D.L.	P.B.C.	D.D.P.
002	.404	.280	.31	059	.371	-.254	-.24
010	.609	-.121	-.15	061	.371	.424	.44
011	.497	.373	.34	062	.629	.388	.42
012	.437	.372	.36	065	.589	.161	.11
014	.497	.334	.30	069	.424	.123	.12
015	.351	.139	.15	071	.450	.315	.34
020	.464	.264	.25	077	.662	.014	-.03
021	.417	.393	.37	079	.417	.216	.19
023	.450	.090	.13	082	.404	.247	.25
024	.636	.309	.47	087	.397	.132	.15
025	.536	.246	.24	092	.616	.012	-.04
027	.510	.326	.33	093	.464	.196	.16
030	.384	-.262	-.27	095	.510	.408	.40
031	.536	.116	.11	099	.397	.140	.16
034	.450	.395	.37	100	.397	.390	.38
035	.556	.344	.32	101	.583	.271	.23
036	.563	.232	.24	102	.510	.309	.31
041	.536	.325	.33	108	.636	.209	.21
047	.457	.123	.12	110	.483	.395	.36
048	.490	.373	.30	114	.543	.252	.30
053	.510	-.015	-.01	126	.351	.193	.13
055	.464	.184	.23	127	.649	-.095	-.13
058	.450	.023	-.02				

4. Items meeting criterion of point biserial correlation.

Revisions needed in content and/or structure to avoid misleading alternatives.

ITEM	D.L.	P.B.C.	D.D.P.
001	.305	.514	.46
004	.748	.431	.38
013	.318	.537	.46
019	.755	.453	.40
123	.722	.491	.51
128	.722	.442	.36

5. Items meeting no criteria. a. difficulty above .650

Revisions needed to content and/or structure to decrease appeal of keyed distractor to non-competent students, and decrease appeal of alternatives to competent students.

ITEM	D.L.	P.B.C.	D.D.P.
005	.834	.267	.19
008	.887	.177	.10
009	.894	.218	.13
029	.934	.269	.10
040	.788	.239	.20
042	.828	.231	.14
049	.715	.324	.32
070	.795	.216	.19
080	.874	.163	.10
081	.662	.316	.34
083	.735	.051	.04
094	.675	.384	.33
098	.828	.254	.17
104	.728	.077	.10
113	.881	.206	.10
117	.742	.421	.40

5. Items meeting no criteria. b. difficulty below .360

Revisions needed to ascertain importance of content, and to increase appeal of keyed distractor to competent students.

ITEM	D.L.	P.B.C.	D.D.P.	ITEM	D.L.	P.B.C.	D.D.P.
003	.238	.294	.28	068	.152	-.073	-.10
006	.305	.274	.22	072	.344	.336	.31
007	.331	.005	-.03	073	.351	.028	.04
016	.285	.192	.20	075	.291	.318	.31
017	.298	.288	.25	078	.179	.369	.31
022	.132	-.134	-.07	084	.311	.416	.40
026	.119	.174	.11	085	.338	.355	.37
028	.093	-.179	-.07	090	.192	-.091	-.07
033	.258	.060	.07	096	.179	.049	.05
038	.245	.186	.16	097	.272	.176	.11
043	.305	.352	.28	103	.252	.360	.28
045	.192	.019	.02	105	.172	.043	.05
046	.185	.331	.19	107	.205	.344	.20
050	.318	.422	.35	109	.245	-.077	-.13
051	.285	.059	.07	111	.252	-.101	-.13
052	.026	-.112	-.03	112	.139	.297	.24
054	.179	.067	.01	115	.358	.421	.34
056	.212	-.069	-.11	116	.311	.429	.32
057	.291	.304	.31	118	.351	.290	.26
060	.106	.076	.05	120	.126	.061	.02
063	.073	-.066	-.03	121	.119	-.185	-.09
064	.152	.201	.17	122	.278	.258	.22
066	.113	.045	.00	124	.338	.287	.20